10. Krise der Prüfungsvalidität

Kapitel 9 hat aufgezeigt, dass traditionelle Lehrformate unter den Bedingungen generativer KI in grundlegende Schwierigkeiten geraten. Für traditionelle Prüfungsformate stellt sich die Lage noch grundlegender dar. Das vorliegende Kapitel widmet sich der Forschungsfrage: Wie können Prüfungen im KI-Zeitalter valide Kompetenzaussagen treffen? Die Antwort erweist sich als schwieriger als die korrespondierende Frage im Bereich der Lehre. Während Lehrveranstaltungen von der Transmission zur Transformation pivotieren und dabei Funktionen bewahren können, die KI nicht zu replizieren vermag, sieht sich das Prüfungswesen mit einem fundamentaleren Problem konfrontiert: KI kann die Produkte erzeugen, die Prüfungen traditionell zu messen beabsichtigten. Eingereichte Artefakte liefern keine belastbaren Evidenz mehr für die Kompetenzen, deren Nachweis es dienen sollte.

Die akademische Bewertungskrise manifestiert sich auf drei Ebenen. Erstens verlieren die dominanten Prüfungsformate — Essays, Hausarbeiten, Aufgabenserien, Abschlussarbeiten — ihre Validität, sobald KI sie kompetent bearbeiten kann. Dies ist kein bloßes Erkennungsproblem, sondern eine grundlegende Erosion des Evidenzverhältnisses zwischen dem beurteilten Produkt und der dahinterstehenden Kompetenz. Zweitens konstituieren viele der vorgeschlagenen Lösungsansätze eine Rückkehr zu Prüfungsformen des 19. Jahrhunderts — beaufsichtigte Klausuren, mündliche Prüfungen, verstärkte Überwachung —, die zwar KI-Nutzung unterbinden, dabei jedoch pädagogische Errungenschaften opfern. Drittens sehen sich zukunftsorientierte Alternativen — prozessbasierte Prüfung, authentische Aufgabenstellungen, KI-integrative Bewertung — mit erheblichen Implementierungshürden konfrontiert und können bestehende Bildungsungleichheiten unter Umständen eher verschärfen als verringern.

Vier analytische Stränge strukturieren die folgende Untersuchung. Abschnitt 10.1 diagnostiziert, auf welchem Wege KI reproduktive Prüfungsformate entwertet, und begründet, warum das Problem über akademische Integritätsfragen hinausreicht. Abschnitt 10.2 analysiert die defensiven institutionellen Reaktionsstrategien — beaufsichtigte Klausuren, technische Überwachungssysteme und mündliche Prüfungen —, die jeweils eine trügerische Sicherheit bieten, dabei aber mit Kosten verbunden sind, die sie als systemische Lösungen erheblich einschränken. Abschnitt 10.3 untersucht KI-resistente Prüfungsansätze — Prozessdokumentation, authentische Aufgaben, kollaborative Formate —, die Validität wahren, dabei aber mit ernsthaften Skalierungs- und Gerechtigkeitsproblemen konfrontiert sind. Abschnitt 10.4 erörtert KI-integrative Prüfung, die kompetente KI-Nutzung als legitimes Bildungsziel begreift, ohne jedoch bereits auf validierte Instrumente oder einen pädagogischen Konsens hinsichtlich der Umsetzung zurückgreifen zu können.

10.1. Die Entwertung reproduktiver Prüfungsformate

10.1.1. Die strukturelle Anfälligkeit schriftlicher Prüfungsformate

Schriftliche Prüfungen dominieren die hochschulische Prüfungspraxis in einem Ausmaß, das ihre historische Bedingtheit verdeckt. Der Essay etablierte sich als Standardprüfungsformat im Zuge der Expansion und Modernisierung der Universitäten im 19. Jahrhundert und verdrängte schrittweise die mündliche Prüfung als primären Mechanismus zur Beurteilung studentischer Leistungen (Broadfoot, 2007). Diese Entwicklung resultierte aus dem Zusammenwirken mehrerer Faktoren: Steigende Studierendenzahlen machten individuelle mündliche Prüfungen praktisch undurchführbar; bürokratische Standardisierungserfordernisse verlangten vergleichbare Beurteilungsmaßstäbe über Studierende und Institutionen hinweg; und die Etablierung von Schriftsprachkompetenz als universellem Bildungsziel positionierte das Schreiben als naheliegende Form des Kompetenznachweises. Bis zur Mitte des 20. Jahrhunderts hatte der Essay eine Dominanz erreicht, die Alternativen als begründungspflichtige Abweichung erscheinen ließ, während er selbst keiner gesonderten Rechtfertigung mehr zu bedürfen schien.

Die Dominanz des Essays beruht auf mehreren impliziten Annahmen, die einer kritischen Reflexion bedürfen. Die erste ist die Schreiben-als-Denken-Annahme: Der Akt des Schreibens offenbare kognitive Prozesse und reflektiere sie, weshalb schriftliche Produkte als verlässliche Indikatoren für das zugrundeliegende Verständnis gelten könnten. Diese Annahme vermengt jedoch Schreibfertigkeit mit konzeptuellem Verständnis und übersieht, dass Studierende Konzepte verstehen können, ohne sie flüssig formulieren zu können — und umgekehrt Gedanken gewandt ausdrücken können, ohne tiefes Verständnis zu besitzen. Die zweite ist die Autorschaftsannahme: Eingereichte

Arbeiten repräsentierten primär die eigene kognitive Leistung der Studierenden, wobei externe Unterstützung auf akzeptable Grenzen beschränkt bleibe. Diese Annahme war stets fragil — Studierende haben seit jeher Unterstützung von Kommilitonen, Tutoren und kommerziellen Diensten in Anspruch genommen —, behielt aber ihre Plausibilität, solange die Inanspruchnahme externer Hilfe mit einem spürbaren Aufwand verbunden war. Die dritte ist die Trennbarkeitsannahme: Fachwissen und Kommunikationsfähigkeit seien separierbare Kompetenzen, die unabhängig voneinander oder kombiniert beurteilbar seien. Diese Annahme rechtfertigt es, Essays als valide Maße disziplinären Verständnisses zu behandeln, obgleich sie zugleich Schreibfähigkeit erfassen und Beurteilungen dadurch potenziell verzerren.

Diese Annahmen finden zusätzliche Stütze in den praktischen Vorzügen des Formats. Schriftliche Prüfungen lassen sich effizient skalieren: Lehrende können eingereichte Arbeiten asynchron bewerten, was die Beurteilung größerer Studierendenkohorten ermöglicht als mündliche Prüfungen. Schriftliche Produkte liefern dokumentarische Evidenz für Qualitätssicherungsverfahren und im Falle von Noteneinsprüchen. Bewertungsraster und standardisierte Kriterien versprechen Objektivität, wenngleich die empirische Forschung konsistent erhebliche Beurteilervarianz nachweist (Boud & Falchikov, 2007). Der kumulative Effekt ist, dass der Essay zum Standardformat der Prüfungspraxis geworden ist, während alternative Formate einer gesonderten Begründung bedürfen.

Die Leistungsfähigkeit generativer KI reicht weit über einfache Textgenerierung hinaus und umfasst die Bewältigung akademischer Aufgaben in unterschiedlichsten Domänen. Zeitgenössische Sprachmodelle demonstrieren Kompetenz in der Abfassung analytischer Essays, der Literatursynthese, der schrittweisen Lösung von Problemen mit begleitenden Erläuterungen, der Code-Generierung mit Dokumentation sowie der Interpretation statistischer Analysen (Mollick & Mollick, 2023). Dieses Fähigkeitsspektrum stellt abgestufte Herausforderungen für verschiedene Prüfungsformate dar, wobei reproduktive Aufgaben — solche, die die Synthese vorhandenen Wissens erfordern, statt genuines neues Wissen zu erzeugen — sich als besonders anfällig erweisen. Ein Sprachmodell mit Zugang zu einschlägigen Texten und theoretischen Konzepten kann Essays generieren, die scheinbares Verständnis demonstrieren, plausible Argumente konstruieren, angemessene Belege anführen und eine kohärente Argumentation aufrechterhalten — genau jene Qualitäten, die Bewertungsraster honorieren. KI-generierte Antworten erreichen folglich häufig die Qualitätsschwelle, die für ein Bestehen erforderlich ist, oder überschreiten diese (Desai, 2025). Frühere KI-Systeme wiesen noch erkennbare Schwächen auf: schematische Struktur, oberflächliche Analyse, gelegentliche Sachfehler. Aktuelle Systeme erzeugen Antworten, die von

Beurteilenden häufig als gleichwertig mit oder überlegen gegenüber durchschnittlicher studentischer Arbeit eingestuft werden (Li et al., 2025).

Selbstauskünfte aus den Jahren 2023 und 2024 zeigen, dass 30 bis 60 Prozent der Studierenden KI-Unterstützung für akademische Arbeiten in Anspruch nehmen, mit erheblicher Variation nach Fach, Institution und Aufgabentyp (Eaton & Turner, 2024). Diese Zahlen dürften den tatsächlichen Nutzungsumfang aufgrund sozialer Erwünschtheit im Antwortverhalten unterschätzen. Mittelbare Hinweise liefern sprunghafte Veränderungen in den Einreichungsmustern: Lehrende berichten von gestiegener Durchschnittsqualität, verringerter Streuung, anspruchsvolleren Formulierungen und charakteristischen Wendungen, die sich über mehrere Einreichungen hinweg wiederholen. Zusammengenommen belegt diese Evidenz, dass die Autorschaftsannahme, auf der schriftliche Prüfungsformate beruhen, empirisch nicht mehr aufrechtzuerhalten ist.

10.1.2. Der Zusammenbruch der Prüfungsvalidität

Die Prüfungskrise berührt im Kern die Frage der Validität — verstanden als den Grad, zu dem Prüfungen tatsächlich messen, was sie zu messen beanspruchen. Eine als Messinstrument für kritisches Denken konzipierte Essayaufgabe verliert ihre Validität, sobald Studierende KI-generierte Texte einreichen können, die scheinbar kritisches Denken demonstrieren, ohne dass die Studierenden selbst kritisch-analytische Denkprozesse durchlaufen hätten. Das eingereichte Artefakt liefert keine belastbare Evidenz mehr für die analytischen Fähigkeiten der Studierenden. Dieses Validitätsproblem ist von Fragen der akademischen Integrität konzeptuell zu trennen: Selbst wenn Studierende KI mit vollständiger Transparenz und institutioneller Genehmigung verwenden, entstehen Validitätsprobleme, sobald die KI-Nutzung grundlegend verändert, was die Prüfung noch zu messen vermag. Die Krise ist folglich nicht primär moralischer Natur — als Problem des Betrugs — zu verstehen, sondern epistemischer: Prüfungen messen nicht mehr das, was sie zu messen beanspruchen.

Das Ausmaß dieses Validitätsproblems reicht weit über textbasierte Fächer hinaus. Im Ingenieurwesen ergab eine institutionenübergreifende Benchmark-Studie an sieben australischen Universitäten, dass KI für die Mehrzahl der Prüfungsaufgaben verwertbare Antworten mit geringem Anpassungsaufwand generieren konnte — darunter schriftliche Aufgaben, Entwurfsberichte und technische Analysen aus zehn ingenieurwissenschaftlichen Fächern (Nikolic et al., 2023). Eine parallele Analyse von Klausuren im Bereich Software-Engineering an einer spanischen Universität zeigte, dass KI Fragen verschiedener Formate und Schwierigkeitsgrade über mehrere Fächer hinweg erfolgreich bearbeiten konnte (López-Fernández & Vergaz, 2024). Im Maschinenbau erzielte KI starke Ergebnisse bei theoriebasierten Fragen, während sie bei komplexen numerischen Berechnungen schwächer abschnitt

— ein Befund, der einen fachinternen Gradienten der Anfälligkeit offenbart: Die schriftlichen, konzeptuellen Anteile ingenieurwissenschaftlicher Prüfungen sind kompromittiert, während berechnungsintensive Aufgaben eine partielle Resistenz bewahren (Frenkel & Emara, 2024).

Mathematik, aufgrund ihrer formalen Strenge häufig als immun betrachtet, zeigt ein differenzierteres Bild. Eine Analyse universitärer Zulassungstests ergab, dass KI Durchschnittsstudierende in Stochastik und Statistik deutlich übertrifft, während die Leistung in Algebra und Analysis weniger konsistent ausfällt (Udias et al., 2024). Auf dem Niveau standardisierter Bachelor-Aufgaben generiert KI hinreichend plausible und bestandsrelevante Antworten auf Wissens- und Verständnisfragen — die den Großteil routinemäßiger Prüfungsleistungen ausmachen —, während Aufgaben mit höheren kognitiven Anforderungen eine gewisse Resistenz behalten (Dao & Le, 2023). Die Implikation ist bedeutsam: Jene Teile mathematischer Prüfungen, die am häufigsten zur Beurteilung des Bachelorstudiums eingesetzt werden, erweisen sich als besonders anfällig für KI-Substitution.

In den Gesundheitswissenschaften — mit Ausnahme der Medizin — ist das Validitätsproblem ebenso ausgeprägt. Eine vergleichende Studie zeigte, dass GPT-4 bei pflegerischen Staatsexamina sowohl in den USA (NCLEX-RN) als auch in China besser abschnitt als frühere Systemversionen und damit demonstrierte, dass wissensbasierte Zulassungsprüfungen in der Pflege zunehmend im Kompetenzbereich aktueller Sprachmodelle liegen (Wu et al., 2024). Eine systematische Übersichtsarbeit und Metaanalyse von 23 Studien zu nationalen Lizenzierungsprüfungen in Pflege, Pharmazie und Zahnmedizin ergab, dass GPT-4 in allen vier Gesundheitsberufen Bestehensleistungen erreichte (Jin et al., 2024). Für die Ausbildung in den Gesundheitswissenschaften bedeutet dies: Schriftliche und wissensbasierte Prüfungen, die das Vorklinikum und die Grundlagenlehre dominieren, können nicht mehr als verlässliche Indikatoren klinischer Handlungskompetenz gelten.

Diese Prüfungskrise trifft leistungsschwächere Studierende überproportional hart (vgl. das in Kapitel 8 beschriebene Novizen-Paradox) und erzeugt Validitätsdefizite genau dort, wo diagnostisches Feedback am dringendsten benötigt würde. Die Konsequenzen für die Leistungsbeurteilung sind entsprechend gravierend. Lehrende, die eingereichte Arbeiten bewerten, stehen vor einer strukturell unlösbaren Aufgabe: Sie sollen zwischen studentischer Eigenleistung und KI-Leistung unterscheiden — oder realistischer: den Grad des KI-Anteils einschätzen. Die Forschung zur KI-Erkennung belegt, dass eine verlässliche Bestimmung grundsätzlich nicht möglich ist (Weber-Wulff et al., 2023). Selbst erfahrene Prüfende können KI-generierten oder KI-assistierten Text nicht konsistent identifizieren. Schwerwiegender noch: Der Versuch, diese Unterscheidung zu treffen, bindet erhebliche Zeit und Energie der Lehrenden, die stattdessen in genuin pädagogische Tätigkeiten investiert werden könnten.

Harari (2011) unterscheidet zwischen objektiven Realitäten (physikalische Gesetze), subjektiven Realitäten (individuelle Empfindungen) und intersubjektiven Realitäten — Phänomenen, die ausschließlich durch kollektive Überzeugung existieren: Geld, Nationen, Menschenrechte. Akademische Noten gehören dieser dritten Kategorie an. Sie erfüllen ihre Funktion nicht, weil sie Kompetenz objektiv messen, sondern weil alle Beteiligten — Studierende, Lehrende, Arbeitgeber — daran glauben, dass eine bestimmte Note erworbenes Wissen signalisiert. Generative KI untergräbt nun die faktische Grundlage dieser intersubjektiven Übereinkunft. Die zuvor bestehende evidentielle Verbindung zwischen Note und kognitiver Leistung ist nicht mehr verifizierbar. Im Unterschied zu objektiven Realitäten löst diese Erosion keinen unmittelbaren Zusammenbruch aus — intersubjektive Realitäten können bemerkenswert lange fortbestehen, selbst während ihre Grundlagen erodieren. Harari verweist auf Fiatwährungen, die ohne Golddeckung funktionieren, solange das Vertrauen anhält.

Hochschulen befinden sich damit in einer prekären Zwischenposition: Sie vergeben Noten weiterhin nach tradierten Mechanismen, obgleich schwerwiegende Validitätsdefizite bestehen. Diese institutionelle Beharrung ist keine Nachlässigkeit, sondern systemische Notwendigkeit. Eine offizielle Anerkennung der Ungültigkeit würde das gesamte Zertifizierungssystem delegitimieren. So persistiert die institutionelle Fiktion — Prüfungsordnungen werden marginal angepasst, KI-Erkennungssysteme eingesetzt, Präsenzklausuren intensiviert —, während die intersubjektive Realität „guter Abschluss gleich kompetente Absolventin bzw. kompetenter Absolvent" stabilisiert werden muss, obwohl ihre epistemische Basis erodiert. Je länger die Lücke zwischen kollektivem Glauben und messbarer Realität wächst, desto schwerwiegender wird die Glaubwürdigkeitskrise ausfallen, wenn das kollektive Vertrauen schließlich bricht.

Die Implikationen reichen über einzelne Lehrveranstaltungen hinaus. Akademische Zeugnisse und Abschlüsse beziehen ihren Wert daraus, dass sie als verifizierte Kompetenznachweise dienen. Arbeitgeber vertrauen darauf, dass Absolventinnen und Absolventen über die Fähigkeiten verfügen, die ihre Noten ausweisen. Professionelle Lizenzierung stützt sich auf Bildungsabschlüsse als Nachweis der Qualifikationsvoraussetzungen. Können Prüfungen nicht mehr verlässlich bestätigen, dass Studierende die ausgewiesenen Kompetenzen tatsächlich erworben haben, gerät das gesamte akademische Zertifizierungssystem unter strukturellen Vertrauensdruck. Die in Abschnitt 8.1 dokumentierte Lücke zwischen der von Arbeitgebern wahrgenommenen Berufsfähigkeit von Absolventinnen und Absolventen und dem tatsächlichen Anforderungsniveau hat strukturelle Veränderungen in der Personalauswahl beschleunigt: Formale Abschlussanforderungen wurden in vielen Branchen stillschweigend zurückgebaut, während kompetenzbasierte Auswahlverfahren rasch an Bedeutung gewonnen haben (Burning Glass Institute, 2022; TestGorilla, 2023).

10.2. Defensive Prüfungsstrategien und ihre Grenzen

10.2.1. Beaufsichtigte Klausur, technische Überwachung und mündliche Prüfung

Angesichts des Zusammenbruchs der Prüfungsvalidität haben Institutionen drei defensive Reaktionsstrategien verfolgt: die Rückkehr zur beaufsichtigten Präsenzklausur, den Einsatz technischer Überwachungssysteme sowie die Wiederbelebung der mündlichen Prüfung. Jede dieser Strategien adressiert den KI-Zugang, nicht aber das zugrundeliegende Problem — und jede ist mit Kosten verbunden, die ihre Tauglichkeit als systemische Antwort auf die Prüfungskrise wesentlich begrenzen.

Beaufsichtigte Präsenzklausuren eliminieren den KI-Zugang durch erzwungene technologische Isolation. Die Logik ist nachvollziehbar, die strukturellen Folgekosten sind jedoch erheblich. Zeitbeschränkungen verhindern komplexe Analysen; geschlossene Formate prämieren Reproduktion gegenüber Anwendung; prüfungsbedingte Belastungssituationen erzeugen Leistungsangst, die gerade jene Studierenden benachteiligt, die am sorgfältigsten denken (Black & Wiliam, 1998). Das Format, das der KI-Druck wiederbelebt, ist dasselbe Format, von dem sich die hochschuldidaktische Entwicklung aus sachlich begründeten Erwägungen entfernt hatte (Broadfoot, 2007).

Überwachungsbasiertes Proctoring versucht, die Flexibilität häuslicher Prüfungssituationen zu erhalten und dabei KI-Nutzung durch Webcam-Überwachung, Bildschirmaufzeichnung und automatisierte Verhaltensanalyse zu unterbinden. Der Ansatz scheitert auf mehreren Ebenen: Technische Umgehungsmöglichkeiten bleiben vergleichsweise unkompliziert, Falsch-Positiv-Raten sind hoch, und die kontinuierliche Aufzeichnung häuslicher Umgebungen wirft ernsthafte datenschutzrechtliche und grundrechtliche Fragen auf (Swauger, 2020; Eaton & Turner, 2024). Über diese praktischen Mängel hinaus entfaltet das Instrument einen strukturellen Folgeschaden: Überwachung signalisiert institutionelles Misstrauen und überführt die Prüfungssituation in ein antagonistisches Verhältnis zwischen Institution und Studierenden. Sowohl die Rückkehr zur Präsenzklausur als auch die Überwachungslösung belasten überproportional jene Studierenden, denen kein ruhiger Arbeitsraum, eine stabile Internetverbindung oder die körperlichen und kognitiven Voraussetzungen fehlen, die Proctoring-Algorithmen als Normalzustand voraussetzen.

Mündliche Prüfungen bieten eine stärkere KI-Resistenz — verbale Interaktion in Echtzeit mit adaptivem Nachfragen lässt sich strukturell kaum auslagern (Joughin, 2010; Hartmann, 2025; Ward et al., 2024). Die Skalierungsgrenzen sind jedoch in vielen Fällen prohibitiv: Individuelle mündliche Prüfungen erfordern 20 bis 30 Minuten je Studierender bzw. Studierendem, was eine vollständige Abdeckung in großen Lehrveranstaltungen ohne erhebliche

Ressourceninvestitionen ausschließt, die die meisten Institutionen nicht dauerhaft aufbringen können.

Alle drei Strategien folgen derselben Grundlogik: den KI-Zugang zu unterbinden, statt das Prüfungswesen unter KI-Bedingungen neu zu konzipieren. Warum diese Logik nicht nur praktisch unzulänglich, sondern analytisch zu kurz greift, zeigt der folgende Abschnitt.

10.2.2. Die Grenzen der Ausschlusslogik

Die zuvor beschriebenen Strategien behandeln KI als externe Störgröße, die es institutionell abzuwehren gilt. Diese Rahmung erweist sich nicht nur als praktisch wenig wirksam — wie die vorliegende Evidenz zeigt —, sondern als analytisch verfehlt. Wenn KI-Werkzeuge in den Berufsfeldern, auf die Studierende sich vorbereiten, bereits zum Standardrepertoire professioneller Praxis gehören, schützen Prüfungssysteme, die KI ausschließen, keine akademische Integrität: Sie produzieren Absolventinnen und Absolventen, deren nachgewiesene Kompetenzen unter Bedingungen erbracht wurden, die außerhalb der Hochschule längst nicht mehr existieren. Die maßgebliche Frage lautet daher nicht, wie KI aus Prüfungen ferngehalten werden kann, sondern wie unter KI-Bedingungen aussagekräftig geprüft werden kann.

Das grundlegende Problem der Ausschlussstrategien liegt darin, dass sie auf die Verhinderung von KI-Nutzung hin optimieren, nicht auf die Förderung von Lernen. Prüfungen erfüllen mehrere Funktionen gleichzeitig: Sie liefern Rückmeldungen, die das weitere Lernen leiten; sie messen Lernergebnisse für Beurteilungszwecke; sie zertifizieren Kompetenzen gegenüber externen Adressaten; und sie motivieren das studentische Engagement mit den Lerninhalten. Wenn KI-Resistenz zum primären Gestaltungskriterium wird, werden diese Bildungsfunktionen den Sicherheitsinteressen untergeordnet. Das Ergebnis sind Prüfungsformate, die Täuschungsversuche unter Umständen wirksam einschränken, dabei jedoch verfehlen, was eigentlich geprüft werden sollte, und Lernen nicht wirksam unterstützen.

Hinzu kommt, dass Ausschlussstrategien die Realität ignorieren, dass KI in den späteren Berufskontexten der Studierenden allgegenwärtig sein wird. Prüfungen, die KI-Nutzung verbieten, erzeugen eine strukturell inkohärente Situation: Das Bildungssystem qualifiziert Studierende für den Verzicht auf Werkzeuge, die sie im Beruf unausweichlich einsetzen werden, und beurteilt ihre Berufsfähigkeit auf der Grundlage von Leistungen unter künstlichen Beschränkungen. Die Diskrepanz zwischen Bildungs- und Berufspraxis wächst genau in dem Moment, in dem strukturelle Kohärenz am dringendsten wäre.

Die Opportunitätskosten dieses Ansatzes verdienen besondere Beachtung. Zeit und Energie, die in KI-Erkennung, Prüfungssicherheit und Integritätsermittlungen fließen, könnten stattdessen in pädagogische

Weiterentwicklung investiert werden. Institutionen, die in Überwachungsinfrastruktur investieren, desinvestieren gleichzeitig in Lehre, Curriculumentwicklung und Studierendenbetreuung (Eaton & Turner, 2024).

Ausschlussstrategien adressieren schließlich auch die in Kapitel 6 analysierte motivationale Krise nicht. Studierende, die KI-Nutzung als legitimes Produktivitätswerkzeug betrachten, erleben institutionelle Verbote als willkürliche Einschränkung statt als begründete Grenzziehung. Diese Wahrnehmung untergräbt die intrinsische Lernmotivation und fördert eine instrumentelle Orientierung: Studierende richten ihre Aufmerksamkeit auf die formale Einhaltung von Regeln statt auf das inhaltliche Engagement mit dem Lernstoff. Die für Ausschlussstrategien erforderlichen Kontrollmaßnahmen signalisieren institutionelles Misstrauen und belasten das pädagogische Verhältnis zwischen Lehrenden und Studierenden. Das Ergebnis ist eine Dynamik wechselseitiger Anpassung — Institutionen verschärfen Kontrollmechanismen, Studierende entwickeln Umgehungsstrategien —, die genuines Lernengagement systematisch verdrängt.

Die entscheidende Frage ist daher nicht, wie Ausschluss wirksamer durchgesetzt werden kann, sondern ob Prüfungen ihre Validität auch ohne ihn bewahren können. Abschnitt 10.3 untersucht die vielversprechendsten Ansätze.

10.3. KI-resistente Prüfung

10.3.1. Zum Begriff der KI-Resistenz

KI-resistente Prüfung bezeichnet Bewertungsformate, in denen KI nicht wirksam als Ersatz für die eigene Leistung der Studierenden eingesetzt werden kann — entweder weil die Prüfung Fähigkeiten verlangt, die KI nicht besitzt, oder weil die Prüfungsstruktur KI-Unterstützung erkennbar oder schlicht irrelevant macht. Der Begriff bedarf einer präzisen Bestimmung. KI-Resistenz unterscheidet sich von bloßer Schwierigkeit der Erkennung: Eine Prüfung ist nicht bereits dann KI-resistent, wenn KI-generierte Antworten schwer zu identifizieren sind. Tatsächliche Resistenz bedeutet, dass der Versuch, KI einzusetzen, entweder keine hinreichenden Antworten erzeugt oder studentische Fähigkeiten erfordert, die einer eigenständigen Bearbeitung funktional gleichwertig sind.

Drei distinkte Resistenzmechanismen lassen sich unterscheiden. Temporale Resistenz beruht auf Zeitbeschränkungen und synchroner Interaktion: mündliche Präsentationen mit spontanem Nachfragen, schriftliche Aufgaben unter Aufsicht, zeitgebundenes Problemlösen. Diese Formate schränken die Möglichkeit zur KI-Nutzung durch ihren Echtzeitcharakter strukturell ein. Interaktive Resistenz erfordert genuine menschliche Interaktion: kollaborative Gruppenarbeit, Debatten mit Peer-Reaktion, Diskussionen, die auf den

Beiträgen anderer aufbauen — Formate, die soziale Aushandlung und adaptives Reagieren auf unvorhersehbares menschliches Verhalten voraussetzen. Kontextuelle Resistenz stützt sich auf lokales oder persönliches Wissen, das für KI nicht zugänglich ist: die Analyse spezifischer Seminardiskussionen, die Reflexion über eigene Felderfahrungen, die Interpretation von Primärquellen außerhalb der KI-Trainingsdaten — Wissen, das ausschließlich in der unmittelbaren Erfahrung der Studierenden verankert ist.

Ein wesentlicher Vorbehalt ist anzubringen: KI-Resistenz ist temporär und graduell, nicht absolut und dauerhaft. Mit der Weiterentwicklung der KI-Fähigkeiten — durch multimodale Eingabe, größere Kontextfenster, agentenbasierte Systeme — dürften gegenwärtige Resistenzmechanismen erodieren. KI-Echtzeit-Assistenz über Audioschnittstellen könnte temporale Resistenz aushöhlen; zunehmend leistungsfähige KI-Agenten mit erweiterter autonomer Handlungsfähigkeit könnten bestimmte kontextuelle Einschränkungen überwinden. Das Ziel besteht daher nicht darin, dauerhaft KI-sichere Prüfungsformate zu entwickeln, sondern Formate zu gestalten, die unter den gegenwärtigen technologischen Bedingungen ihre Validität bewahren und zugleich pädagogisch sinnvoll bleiben — unabhängig von der künftigen Entwicklung der KI.

10.3.2. Prozessbasierte Prüfung

Prozessbasierte Prüfung verlagert den Bewertungsfokus vom Endprodukt auf den Entstehungsprozess. Anstatt lediglich den eingereichten schriftlichen Ausarbeiteungen zu bewerten, werden der Erarbeitungsprozess und seine Dokumentation in die Beurteilung einbezogen: Recherchemitschriften, Gliederungsentwürfe, Zwischenfassungen, Revisionsentscheidungen und Reflexionen über Überarbeitungsschritte. Die theoretische Grundlage entstammt der konstruktivistischen Lerntheorie (vgl. Kapitel 4.2): Lernen vollzieht sich im Prozess des Verstehensaufbaus, nicht allein in den abschließenden Produkten, die daraus resultieren. Die Beurteilung des Prozesses richtet die Prüfung enger an dem aus, was tatsächlich gelernt wird.

Die Umsetzungsformen variieren in Komplexität und Eingriffstiefe. In einer minimalen Variante reichen Studierende Prozessunterlagen zusammen mit dem Endprodukt ein: Recherchenotizen, Gliederungen, Entwurfsfassungen. Lehrende sichten diese Materialien auf Anzeichen vertieften inhaltlichen Engagements. Anspruchsvollere Ansätze nutzen die automatische Versionsverfolgung: Werkzeuge wie Google Docs protokollieren alle Änderungen mit Zeitstempel und erzeugen so eine detaillierte Aufzeichnung des Schreibprozesses einschließlich Pausen, Lösch- und Überarbeitungsmustern sowie Schreibgeschwindigkeit. Fortgeschrittene Varianten erfordern Prozessreflexionen, in denen Studierende ihre

Entscheidungen begründen, aufgetretene Schwierigkeiten dokumentieren und Überarbeitungsschritte erläutern.

Diese Dokumentationsanforderungen entfalten eine besondere diagnostische Prüfkraft: Eine über Wochen konsistent durchgehaltene Argumentation ist praktisch nicht aufrechtzuerhalten, wenn das Verständnis oberflächlich geblieben ist. Studierende, die KI einsetzen, ohne sich inhaltlich vertieft mit dem Stoff auseinanderzusetzen, erzeugen charakteristische Dokumentationsmuster — plötzliche Einsichten ohne stützende Recherche, Argumentationssprünge ohne Zwischenschritte, Richtungswechsel ohne nachvollziehbare Begründung —, die das Fehlen genuiner inhaltlicher Durchdringung offenbaren. Die Anforderung, den Arbeitsprozess sichtbar zu machen, fördert zudem einen systematischeren Ansatz bei der Aufgabenbearbeitung mit expliziter Planung und iterativer Überarbeitung. Prozessbasierte Prüfung erfüllt damit eine doppelte Funktion: Sie erzeugt KI-Resistenz und fördert zugleich Arbeitsweisen, die das Lernen unabhängig von KI-bezogenen Überlegungen verbessern.

Das Format prozessbasierte Prüfung begegnet jedoch erheblichen Grenzen. Zunächst steigt die Arbeitsbelastung für Studierende und Lehrende erheblich: Studierende müssen Tätigkeiten dokumentieren, die typischerweise unsichtbar bleiben, was über die eigentliche Aufgabe hinausgehenden Zeitaufwand erzeugt; Lehrende müssen nicht nur Endprodukte, sondern umfangreiche Prozessunterlagen sichten, was den Beurteilungsaufwand vervielfacht. Darüber hinaus kann die Prozessdokumentation selbst gefälscht werden: Studierende können plausibel wirkende Recherchenotizen generieren, fingierte Entwurfsfassungen erstellen oder fiktive Überarbeitungsnarrative konstruieren — KI kann dabei überzeugende Prozessartefakte produzieren, die möglicherweise authentisch erstellten überlegen sind. Schließlich können Dokumentationsanforderungen neue Formen struktureller Benachteiligung erzeugen: Studierende mit ausgeprägten organisatorischen Fähigkeiten, effektivem Zeitmanagement und Vertrautheit mit akademischen Konventionen bewältigen Prozessdokumentation leichter als Studierende, denen diese Voraussetzungen fehlen.

Die Überwachungsdimension wirft zusätzliche Bedenken auf. Detailliertes Prozessmonitoring — insbesondere automatisierte Versionsverlaufsanalyse — nähert sich einem Maß an Steuerung, das Lernprozesse eher hemmen als fördern kann. Studierende können Prozessanforderungen als invasive Kontrolle ihrer Arbeitsgewohnheiten erleben statt als unterstützende Strukturierungshilfe. Die Forderung, alle Denkschritte sichtbar zu machen, widerspricht dem natürlichen Verlauf kreativer Prozesse, der häufig scheinbar unproduktive Erkundungen, Sackgassen und nichtlineare Entwicklungen einschließt. Prozessbasierte Prüfung riskiert damit, bestimmte Arbeitsstile zu begünstigen und Ansätze zu benachteiligen, die auf weniger konventionellen Wegen zu guten Ergebnissen gelangen.

10.3.3. Authentische Prüfung

Authentische Prüfung verlangt von Studierenden, Aufgaben zu bewältigen, die reale berufliche oder gesellschaftliche Kontexte widerspiegeln, statt Aufgaben zu bearbeiten, die ausschließlich für Zwecke der Leistungsbeurteilung konstruiert wurden (Wiggins & McTighe, 2005). Dem Ansatz liegt die Annahme zugrunde, dass praxisverankerte Aufgaben Fähigkeiten präziser messen und die Lernmotivation nachhaltiger fördern. Authentische Aufgaben integrieren typischerweise mehrere Kompetenzen, erfordern Urteilsvermögen in mehrdeutigen Situationen und erzeugen Ergebnisse, die über den Erwerb einer Note hinaus einen eigenständigen Zweck erfüllen.

Disziplinäre Beispiele verdeutlichen das Konzept. In MINT-Fächern umfasst authentische Prüfung Laborpraktika, in denen Studierende Experimente selbst entwerfen und durchführen, reale Daten analysieren, spezialisierte Geräte bedienen und Probleme beheben müssen — Tätigkeiten, die physische Präsenz und praktische Handlungsfähigkeit erfordern, die KI nicht ersetzen kann. In den Sozialwissenschaften können authentische Aufgaben die eigenständige Durchführung und Auswertung von Interviews, ethnografische Beobachtungen oder die Analyse archivalischer Materialien umfassen, die nicht in digitaler Form vorliegen. In den Geisteswissenschaften entsteht Authentizität durch die Arbeit mit Primärquellen: Manuskriptanalyse, Textkritik oder die Interpretation von Artefakten, die den direkten Umgang mit dem Material erfordern. Professionelle Studiengänge verfügen über strukturell verankerte authentische Prüfungsformen: klinische Rotationen im Medizinstudium, Entwurfsstudios in der Architektur, Unterrichtspraktika in der Lehramtsausbildung.

Authentische Prüfung sieht sich jedoch erheblichen praktischen Einschränkungen gegenüber. Ressourcenintensität bildet die primäre Grenze: Authentische Aufgaben erfordern typischerweise Geräte, Materialien, Betreuungskapazitäten oder Feldzugänge, die kostspielig und begrenzt verfügbar sind. Standardisierung wird schwierig, wenn Studierende mit unterschiedlichen Materialien oder Kontexten arbeiten, was die Vergleichbarkeit der Beurteilung erschwert. Bewertungen werden zeitaufwändiger und stärker von subjektivem Urteil abhängig, je komplexer und kontextgebundener die Aufgaben werden. Vor allem aber lassen sich nicht alle relevanten Lernziele in authentische Aufgabenstellungen überführen: Abstraktes theoretisches Verständnis, grundlegende Fertigkeiten und thematische Wissensbreite besitzen häufig keine unmittelbar entsprechende authentische Prüfungsform jenseits der beruflichen Praxis selbst.

Die in Kapitel 7 eingeführte disziplinäre Asymmetrie tritt hier besonders deutlich hervor. MINT-Fächer verfügen über strukturell verankerte authentische Prüfungsformen: Laborarbeit, Experimentaldesign und instrumentelle Praxis sind reguläre Bestandteile des Curriculums. Professionelle Studiengänge integrieren praxisbasierte Prüfung als

Kernanforderung. Geistes- und Sozialwissenschaften haben demgegenüber größere Schwierigkeiten, authentische Aufgaben zu identifizieren, die skalierbar sind. Diese fachliche Ungleichverteilung bedeutet, dass KI-resistente authentische Prüfung in einigen Disziplinen strukturell leichter zugänglich ist als in anderen — mit dem Risiko, bestehende Bildungsungleichheiten zu verstärken.

10.3.4. Kollaborative und präsentationsbasierte Prüfung

Kollaborative Prüfungsformate — Gruppenprojekte, Teampräsentationen, gemeinsames Problemlösen — erzeugen KI-Resistenz dadurch, dass sie Koordination und Aushandlung zwischen menschlichen Beteiligten erfordern. KI kann zwar zu Gruppenarbeit beitragen, indem sie Inhalte generiert oder Probleme analysiert, nicht jedoch die interpersonale Dynamik ersetzen, die echte Zusammenarbeit konstituiert: Rollen aushandeln, Meinungsverschiedenheiten klären, auf die Beiträge anderer eingehen, Ideen gemeinsam weiterentwickeln (Evans, 2020). Die menschliche Interaktionsdimension setzt der KI-Substitution strukturelle Grenzen.

Präsentationsbasierte Prüfungsformen — mündliche Präsentationen mit Fragerunde, Posterpräsentationen, Debatten — erzeugen temporale und interaktive Resistenz. Studierende müssen ihr Verständnis spontan artikulieren, auf unvorhersehbare Fragen reagieren und ihre inhaltliche Durchdringung durch Erläuterungen in Echtzeit belegen. Während KI für die Vorbereitung eingesetzt werden kann, setzt die Präsentation selbst voraus, dass die Inhalte hinreichend verinnerlicht wurden, um ein freies Gespräch zu tragen. Das Format differenziert damit zwischen Studierenden, die den Stoff verstanden haben, und solchen, die ihn auswendig gelernt oder Informationen ohne tieferes Verständnis abgerufen haben.

Diese Formate sind gleichwohl mit bekannten Grenzen konfrontiert. Gruppenarbeit ist anfällig für das Trittbrettfahrer-Problem: Einzelne Studierende tragen minimal bei, erhalten aber die Bewertung für die Gruppenleistung. Mechanismen zur individuellen Rechenschaftspflicht — Peer-Evaluationen, individuelle Anteile innerhalb von Gruppenarbeiten — helfen, fügen aber Komplexität hinzu. Präsentationsfähigkeit korreliert nicht notwendig mit tiefem Verständnis: Verbal gewandte Studierende können Inhalte überzeugend vortragen, ohne sie inhaltlich durchdrungen zu haben, während fachlich kompetente Studierende mit schwächeren Präsentationsfähigkeiten weniger kompetent erscheinen mögen. Prüfungsangst ist ein realer Faktor: Redeangst beeinträchtigt viele Studierende und kann die Demonstration tatsächlichen Wissens überlagern. Kulturelle und sprachliche Faktoren spielen ebenfalls eine Rolle: Präsentationserwartungen variieren kulturell, und Studierende, die nicht in ihrer Muttersprache präsentieren, sind unabhängig von ihrem konzeptuellen Verständnis strukturell benachteiligt.

10.4. KI-integrative Prüfung

KI-integrative Prüfung beurteilt nicht, ob Studierende KI genutzt haben, sondern ob sie es kompetent getan haben. Diese Unterscheidung verändert den Gegenstand der Beurteilung grundlegend: Anstelle der Aufgabenerledigung rücken die metakognitiven und evaluativen Fähigkeiten in den Mittelpunkt, die eine kompetente KI-Nutzung erfordert — das Formulieren wirkungsvoller Eingabeaufforderungen, die kritische Beurteilung von Ausgaben, das Erkennen von Situationen, in denen menschliches Urteilsvermögen algorithmischen Empfehlungen überlegen ist, und die Integration von KI-Beiträgen in eine kohärente eigenständige Argumentation. Genau diese Fähigkeiten sind es, die berufliche Kontexte zunehmend verlangen — und sie lassen sich nicht durch Verbote entwickeln.

Aus dieser Neuausrichtung folgen drei Gestaltungsanforderungen. *Erstens* ersetzen explizite Nutzungsprotokolle vage Erlaubnisformulierungen. „KI verantwortungsvoll nutzen" ist keine Orientierung, sondern eine Mehrdeutigkeit, die Studierende und Lehrende unterschiedlich und inkonsistent auslegen werden. Wirksame Protokolle legen konkret fest, was KI leisten darf und was Studierende selbst erbringen müssen: Literatur zu identifizieren unterscheidet sich vom Konstruieren von Argumenten; Grammatik zu prüfen unterscheidet sich vom Generieren von Analysen. Die Grenze zwischen unterstützender und substitutiver Nutzung muss hinreichend konkret sein, um lehrbar und beurteilbar zu sein. *Zweitens* macht Prozessdokumentation die KI-Nutzung sichtbar und beurteilbar. Studierende halten fest, welche Eingabeaufforderungen sie formuliert haben, welche Ausgaben sie erhalten haben und wie sie diese bewertet, verändert oder verworfen haben. Die Dokumentation erfüllt dabei zwei Funktionen zugleich: Sie schafft die Evidenzbasis für die Beurteilung der KI-Nutzungsqualität und entwickelt jene metakognitiven Gewohnheiten, die genuine KI-Kompetenz ausmachen. *Drittens* richten Beurteilungskriterien sich an Kompetenz statt an Regelkonformität aus. Maßgeblich ist, ob Studierende Eingabeaufforderungen formuliert haben, die konzeptuelles Verständnis erkennen lassen; ob sie identifiziert haben, wo KI-Ausgaben oberflächlich, unzutreffend oder unzureichend kontextualisiert waren; und ob sie dort eigenständiges Urteilsvermögen eingesetzt haben, wo die Aufgabe es erforderte. Diese Kriterien verlagern die Beurteilung von der Erkennung zur Evaluation.

KI-integrative Prüfung lässt sich nicht durch isolierte Änderungen an traditionellen Formaten umsetzen — etwa durch das nachträgliche Hinzufügen von Offenlegungspflichten zu bestehenden Essays oder durch die bloße Genehmigung von KI-Nutzung mit Quellenangabe. Der Ansatz erfordert eine koordinierte Neugestaltung von Lehrmethoden, Lernstrukturen und Prüfungspraxis, weil die zu beurteilenden Kompetenzen durch kontinuierliche Übung entwickelt werden und nicht in einzelnen Prüfungsereignissen mit hohem Gewicht demonstriert werden können. Dies ist keine Einschränkung

des Ansatzes, sondern eine Klärung seiner Voraussetzungen: Eine Prüfung, die KI-Nutzungskompetenz bewertet, muss in pädagogische Kontexte eingebettet sein, in denen KI-Nutzung praktiziert, reflektiert und schrittweise verfeinert wird.

Kapitel 14 entwickelt ein konkretes Modell für diese Integration im Rahmen des KI-integrierten problembasierten Lernens und zeigt, wie Portfoliodokumentation, explizite Nutzungsprotokolle und mündliche Prüfung zu einer kohärenten Prüfungsarchitektur zusammengefügt werden können — einer Architektur, die sowohl die Grundlagenkompetenzen adressiert, die KI nicht ersetzen kann, als auch die weiterführenden Fähigkeiten, die KI-durchdrungene berufliche Kontexte verlangen.