Frederik von der Heyden

Praxisbooklet / 26 Seiten

Lohnt sich der **KI-Pilot.**

KI

Nutzen messen, bevor du ausweitest

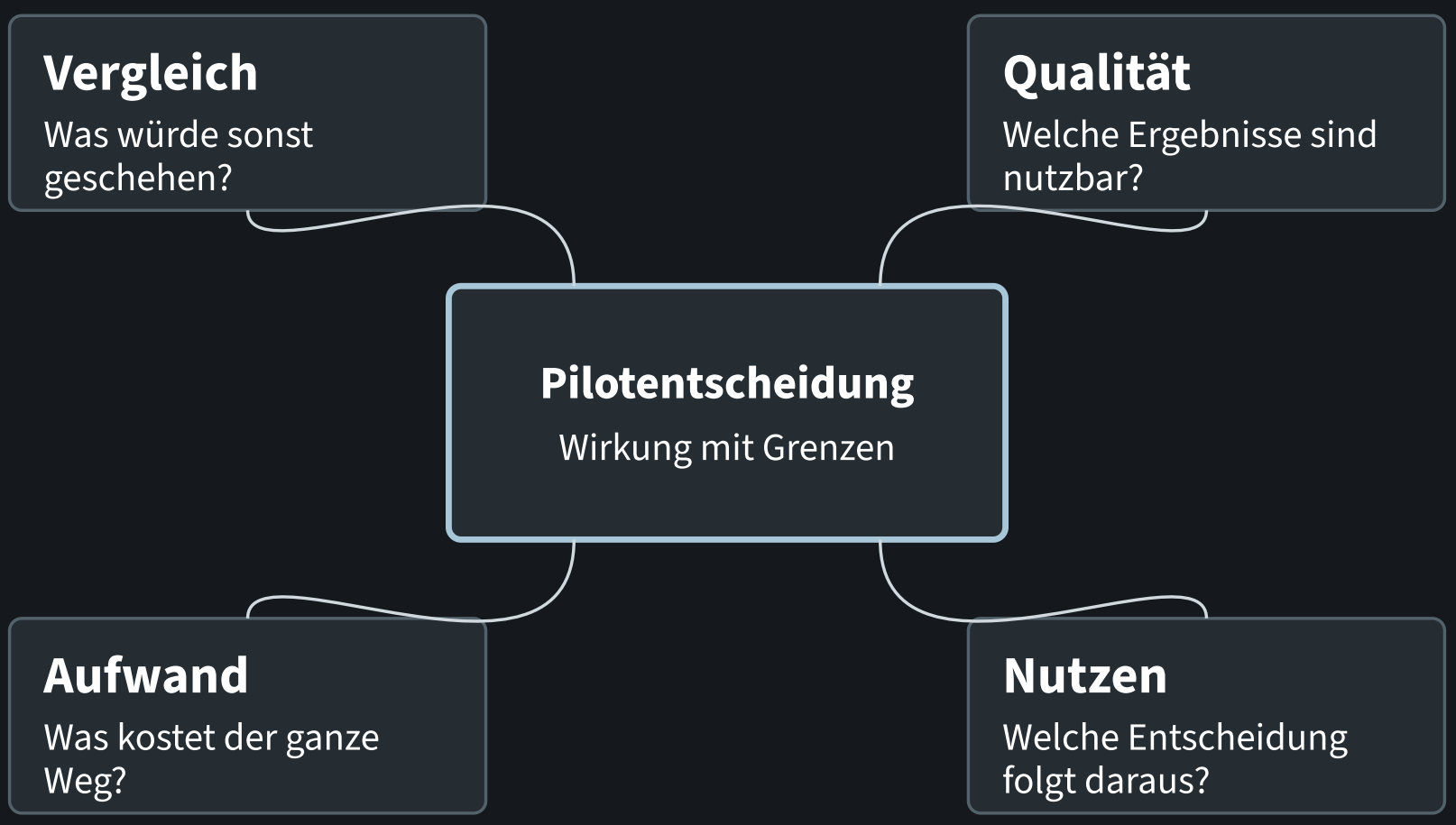
Vergleiche den vollständigen Arbeitsaufwand und trenne mögliche Entlastung von einer tatsächlich belegten Geldwirkung.

KI-gestaltetes persönliches Motiv auf Basis des freigegebenen Bildsatzes.



Eine Entscheidung braucht **mehr als Nutzung**

Nach der Lektüre kannst du einen Pilot fortsetzen, gezielt verändern oder mit nachvollziehbarem Grund beenden.



Viele erzeugte Entwürfe zeigen zunächst nur, dass eine Funktion verwendet wird. Für unsere Entscheidung müssen wir den Ablauf vergleichen, die Qualität prüfen und den verbleibenden Aufwand zählen. Erst danach überlegen wir, ob die beobachtete Entlastung im Team tatsächlich nutzbar wird. Jeder dieser Schritte beantwortet eine andere Frage.

HM Treasury empfiehlt, Evaluation früh in die Gestaltung von KI-Vorhaben einzubauen und die erwartete Wirkung ausdrücklich zu beschreiben. Wir übertragen dieses Grundprinzip auf einen kleinen betrieblichen Pilot. Die Zahlen und der gesamte Versuchsplan sind eigene, konstruierte Lehrbeispiele und keine berichtete Wirkung eines Produkts.

Unser Team **bearbeitet interne Anfragen**

Der fiktive Pilot erstellt Antwortentwürfe zu häufigen Fragen über interne Raum- und Materialbuchungen.

Aufgabe	Interne Antwort vorbereiten und freigeben
Pilot	Suche und erster Entwurf mit KI
Menschlich	Quelle prüfen und Antwort freigeben
Kennzeichnung	Designstudie · kein Kundenauftrag · neutrale Demo-Daten

Bisher sucht eine Servicekraft die Regel nach, formuliert eine Antwort und prüft sie vor der internen Freigabe. Im Pilot übernimmt ein Assistent Suche und ersten Entwurf. Die Servicekraft kontrolliert Quelle, Vollständigkeit und Ton weiterhin selbst. Änderungen an Buchungen sind nicht erlaubt. So bleibt die untersuchte Aufgabe überschaubar.

Wir rechnen mit zwanzig konstruierten Anfragen je Vergleichsweg. Es handelt sich nicht um einen echten Versuch und nicht um eine empfohlene Stichprobengröße. Die Menge dient ausschließlich dazu, Zeitaufwand und Nacharbeit vollständig zeigen zu können. Alle Regeln, Fälle und Beobachtungen sind neutrale Demo-Daten.

Schreibe die Wirkung **als überprüfbare Erwartung**

Die Erwartung nennt den vermuteten Mechanismus und die Bedingung, unter der er nützen würde.

Vermutete Entlastung	Weniger Suche und erster Schreibaufwand
Bedingung	Kontrolle und Nacharbeit bleiben vertretbar
Qualität	Freigegebene Antwort ist fachlich passend
Alternative	Bessere Suche ohne Generierung

Unser Team vermutet, dass ein passender Entwurf Such- und Schreibzeit verringert. Dieser Vorteil wäre nur dann brauchbar, wenn Kontrolle und Korrektur ihn nicht wieder aufzehren und die freigegebenen Antworten fachlich passen. Deshalb messen wir den vollständigen Weg bis zur Freigabe und prüfen die Antworten nach denselben Kriterien.

Die Erwartung ist keine Zusage des Anbieters und keine Erfolgsmeldung. Sie kann durch die Beobachtung widerlegt werden. Halte vor dem Start auch eine Alternative fest, beispielsweise eine bessere Suchmaske. Sonst könnte ein Pilot gewinnen, obwohl eine einfachere Änderung denselben Engpass mit weniger laufendem Aufwand lösen würde.



KI-gestaltetes persönliches Motiv. Keine Aufnahme eines realen Kundenprojekts.

Kapitel 1

01

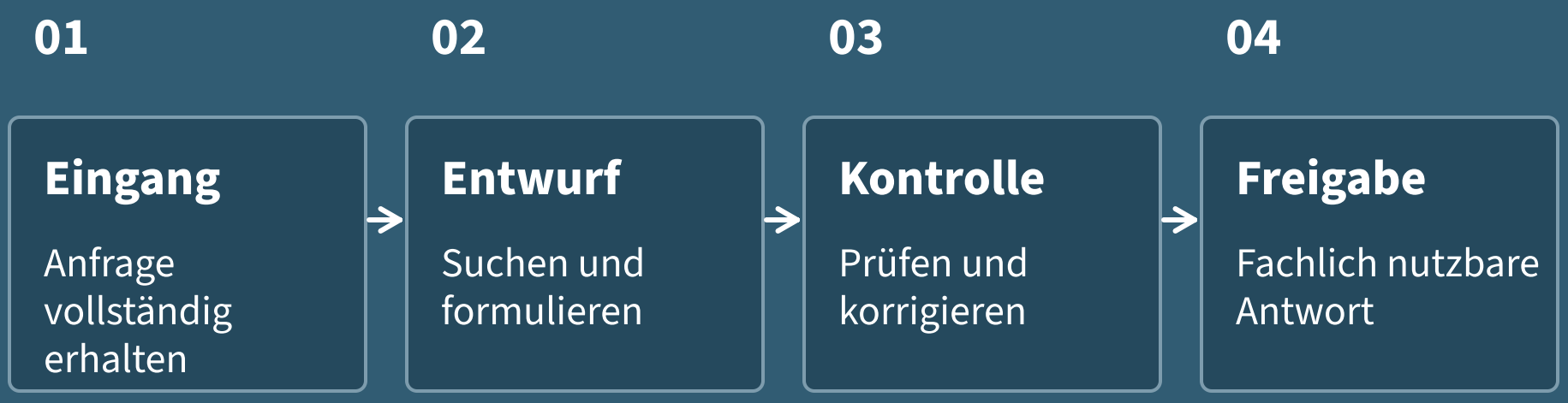
Vergleiche **den vollständigen Arbeitsweg**

Eine schnelle erste Antwort kann den späteren Aufwand unsichtbar machen. Unser Vergleich beginnt deshalb beim Eingang der Anfrage und endet erst bei einer fachlich freigegebenen internen Antwort.

Du legst die Vergleichsbedingungen vor dem Versuch fest und prüfst danach die konstruierten Zeitdaten. Dabei bleiben Qualität, Arbeitskapazität und mögliche Geldwirkung getrennt, damit keine Rechnung mehr verspricht, als sie belegen kann.

Der Entwurf ist **nur eine Zwischenstation**

Der messbare Arbeitsweg umfasst Suche, Erstellung, Kontrolle und die fachlich nutzbare Ausgabe.



Unser Pilot darf nicht bei der ersten generierten Antwort stoppen. Die Servicekraft muss eine falsche Fundstelle berichtigen und eine unklare Regel nachschlagen. Auch diese Minuten gehören zum Vorgang. Wartet sie während einer technischen Störung und kann keine andere Arbeit übernehmen, wird diese gebundene Zeit ebenfalls dokumentiert.

Wartezeit und aktive Arbeitszeit sind dennoch unterschiedliche Größen. Wenn eine Anfrage im Hintergrund verarbeitet wird, kann die Durchlaufzeit steigen, ohne dass dieselbe Person dauerhaft gebunden ist. Lege deshalb fest, welche Zeitart du misst und wofür du sie später verwendest. Eine unklare Mischung erlaubt keine saubere Nutzenrechnung.

Miss den bisherigen Weg **unter denselben Regeln**

Die Ausgangsmessung braucht dieselbe Aufgabe, dieselbe Qualitätsanforderung und einen nachvollziehbaren Zeitraum.

Demomenge	20 vollständig bearbeitete Anfragen
Zeitart	Aktive Arbeit bis zur Freigabe
Demoaufwand	240 Minuten einschließlich Kontrolle
Grenze	Kein gemessener Branchen- oder Kundenwert

In unserer Demo benötigt der bisherige Weg insgesamt 240 aktive Minuten für zwanzig Anfragen. Darin sind Suche, Schreiben und Kontrolle enthalten. Dieser konstruierte Wert ist die Grundlage der späteren Rechnung. Er sagt nichts über das typische Tempo realer Servicearbeit aus und darf nicht als Branchenkennzahl weitergegeben werden.

Für einen echten Versuch dokumentierst du, welche Fälle und Personen betrachtet wurden und ob außergewöhnliche Bedingungen vorlagen. Eine besonders ruhige Woche ist möglicherweise kein guter Vergleich für einen stark belasteten Monat. Eine Ausgangsmessung ist nur so brauchbar wie ihre Übereinstimmung mit der späteren Entscheidung.

Wähle einen Vergleich, der **die Frage trägt**

Ein Vorher-Nachher-Vergleich zeigt eine Veränderung, aber nicht automatisch ihre Ursache.

Vorher und nachher	Einfach, aber zeitliche Einflüsse bleiben
Parallel	Ähnliche Fälle im gleichen Zeitraum
Zuteilung geplant	Vergleichsbedingungen vorab beschreiben
Unsere Demo	Rechenbeispiel ohne kausale Wirkungsaussage

Vielleicht werden nach der Einführung einfachere Anfragen gestellt, vielleicht kennt das Team die Regeln inzwischen besser. Ein paralleler Vergleich unter passenden Bedingungen kann solche Einflüsse besser sichtbar machen. HM Treasurys Magenta Book unterscheidet beobachtete Ergebnisse und die Frage, inwieweit sie der untersuchten Maßnahme zugerechnet werden können.

Unser Lehrfall rechnet zunächst nur zwei beschriebene Wege. Er beansprucht keinen kausalen Nachweis. Wenn deine Entscheidung größere Investitionen oder erhebliche Folgen auslöst, plane eine angemessene Auswertung mit methodischer Unterstützung. Die praktische Form des Vergleichs muss zur Fragestellung und zu den verfügbaren Fällen passen.

Verteile Fälle **vor der Bearbeitung**

Die Fallauswahl darf nicht nachträglich so verändert werden, dass der Pilot besser aussieht.

Zugelassen	Definierte interne Buchungsfragen
Vorab festlegen	Zuteilung, Zeitraum und Zeitart
Mögliche Verzerrung	Erinnerung, Übung und Fallunterschiede
Aufbewahren	Abbrüche und Ausschlussgründe

Im echten Vergleich würden die zugelassenen Anfragearten vorab feststehen. Anschließend würde eine passende Zuteilung geplant und dokumentiert. Wenn Menschen dieselbe Anfrage zweimal bearbeiten, können Erinnerung und Übung den zweiten Weg erleichtern. Solche Einflüsse müssen in der Planung berücksichtigt werden. Unsere konstruierten Zahlen lösen dieses Problem nicht.

Halte außerdem fest, welche Fälle ausgeschlossen werden und warum. Ein Ausschluss nach einem technischen Fehler darf nicht heimlich aus der Auswertung verschwinden. Vielleicht gehört gerade dieser Fehler zum tatsächlichen Pilotaufwand. Ein transparenter Vermerk ist nützlicher als ein scheinbar sauberer Datensatz mit unsichtbarer Auswahl.

Qualität bleibt **eine eigene Bedingung**

Eine Zeitersparnis kann eine fachlich ungeeignete Antwort nicht ausgleichen.

Quelle	Passende Regel und nachvollziehbare Stelle
Inhalt	Vollständige und eindeutige Antwort
Grenze	Keine erfundene Buchungsfreigabe
Entscheidung	Kritischer Fehler verhindert Demo-Freigabe

Unsere Servicekraft prüft Regelbezug, Vollständigkeit und eindeutige Aussage. Ein kritischer Fehler wäre eine erfundene Buchungsfreigabe. Solch eine Antwort würde im Demofall nicht freigegeben, auch wenn sie besonders schnell entstand. Die endgültige Beurteilung liegt bei der fachlichen Rolle und folgt einem vorher beschriebenen Maßstab.

NIST MEASURE 1.1 empfiehlt zum Zweck passende Messgrößen und dokumentierte Grenzen des Messbaren. Wir verwenden deshalb getrennte Angaben für Bearbeitungszeit und fachliche Abweichungen. Eine gemeinsame Gesamtnote würde den Zielkonflikt verdecken. Welche Abweichung in deinem Einsatz die Fortsetzung ausschließt, muss die Verantwortung vor dem Versuch klären.

Unsere **vollständigen Demo-Zeitsummen**

Die folgende Aufteilung umfasst alle zwanzig Anfragen und trennt Entwurf, Kontrolle und Korrektur.

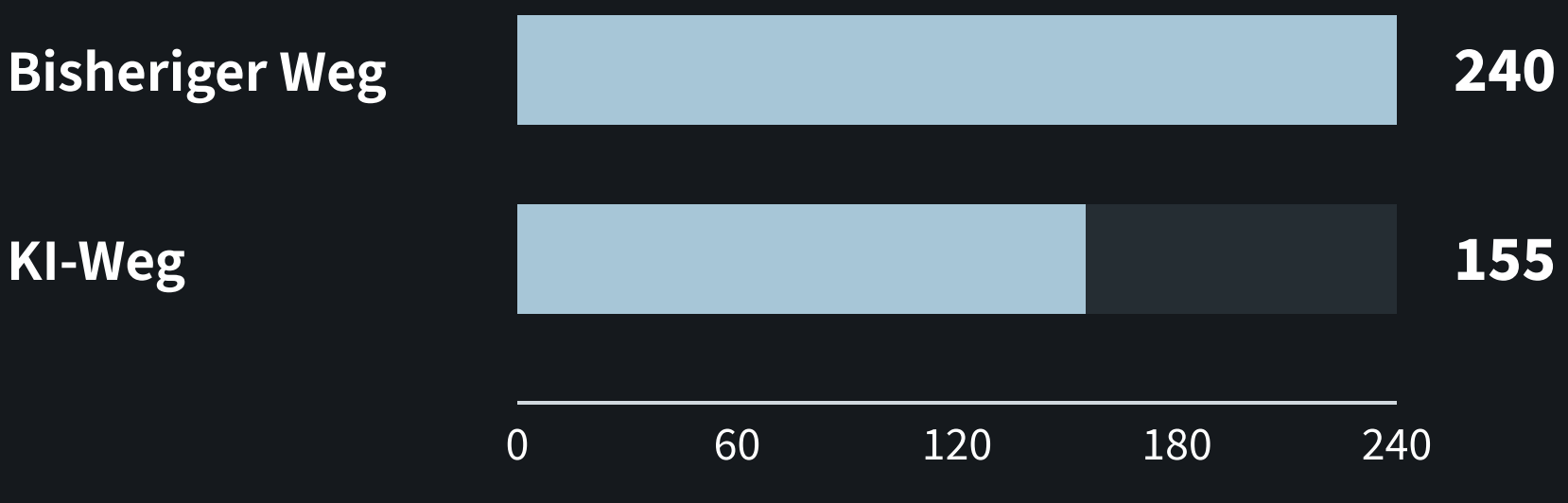
Bisheriger Weg	240 Minuten insgesamt
KI-Entwurfsarbeit	50 aktive Minuten
KI-Kontrolle	80 aktive Minuten
KI-Korrekturen	25 aktive Minuten
KI-Weg insgesamt	155 aktive Minuten

Im KI-Weg fallen 50 Minuten für aktive Entwurfsarbeit, 80 Minuten für Kontrolle und 25 Minuten für Korrekturen an. Zusammen sind das 155 aktive Minuten. Die automatische Hintergrundlaufzeit ist darin nicht als menschliche Arbeit gezählt. Eine echte Aufzeichnung müsste mögliche gebundene Wartezeit gesondert ausweisen.

Die Zahl 155 beschreibt allein unseren konstruierten Datensatz. Sie wurde nicht durch ein Modell oder eine Stoppuhr gemessen. Wir verwenden sie, damit du die Rechnung nachprüfen und die enthaltene Nacharbeit sehen kannst. Die Ausgangssumme 240 enthält die entsprechenden Schritte bereits vollständig und wird nicht noch einmal um Kontrolle ergänzt.

Rechne den **Unterschied pro Fall**

Für zwanzig Demo-Anfragen entstehen 85 Minuten Unterschied zwischen den beschriebenen Arbeitswegen.



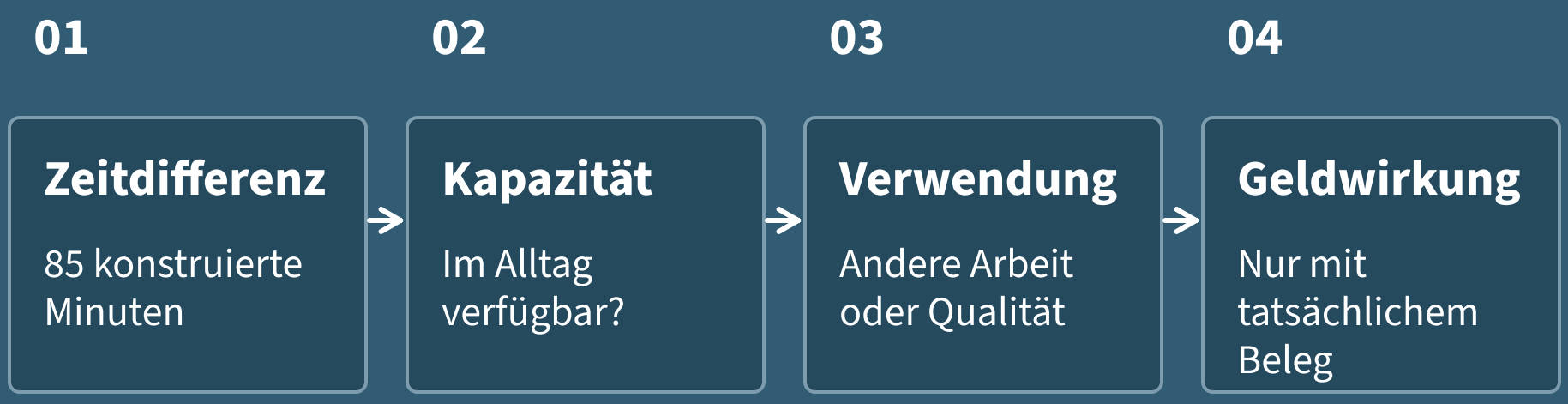
Aktive Arbeit für jeweils 20 Anfragen einschließlich Kontrolle und Korrektur. Konstruierte Demo-Daten; keine Messung.

Die Rechnung lautet 240 minus 155 gleich 85 Minuten. Geteilt durch zwanzig Anfragen ergibt das 4,25 Minuten je Anfrage. Der bisherige Weg liegt bei zwölf Minuten, der KI-Weg bei 7,75 Minuten pro Fall. Diese Mittelwerte erlauben den Vergleich der beschriebenen Gesamtsummen, zeigen jedoch noch nicht ihre Streuung.

Wenn wenige sehr lange Fälle vorkommen, kann derselbe Durchschnitt einen unruhigen Alltag verdecken. Bewahre deshalb für einen echten Pilot die einzelnen Fallzeiten auf und betrachte langsame oder fehlerhafte Fälle gesondert. Die Rechenhilfe darf die Fälle verdichten, aber ihre schwierigen Stellen nicht unsichtbar machen.

Zeitgewinn wird erst **durch Nutzung wirksam**

Freie Minuten sind eine mögliche Kapazität und noch keine realisierte Geldersparnis.



Unsere 85 Demo-Minuten könnten für andere Aufgaben eingesetzt werden. Vielleicht liegen sie jedoch in kleinen Lücken zwischen Anfragen und lassen sich kaum bündeln. Vielleicht werden sie bewusst für eine bessere Prüfung verwendet. Ohne eine tatsächliche Veränderung von Kosten, Kapazität oder Ergebnis bleibt die finanzielle Wirkung offen.

Die Grafik zeigt deshalb getrennte Übergänge. Miss zunächst den vollständigen Aufwand, prüfe dann die nutzbare Entlastung und dokumentiere erst anschließend die tatsächliche Verwendung. Eine angenommene Stundenbewertung wäre eine Planungsannahme. Sie darf nicht als eingesparter Rechnungsbetrag oder realisierter Gewinn bezeichnet werden.

Nimm laufende Pilotarbeit **in die Rechnung**

Der Vorgang kann schneller werden, während Betreuung und Pflege zusätzlichen Aufwand verursachen.

Vorgangsdifferenz	85 Demo-Minuten
Zusätzliche Betreuung	30 Demo-Minuten im gleichen Zeitraum
Mögliche Nettokapazität	55 Demo-Minuten
Noch getrennt	Einrichtung und tatsächliche externe Ausgaben

Für unsere Demo nehmen wir 30 Minuten Betreuung im selben Zeitraum an. Diese Zahl ist ebenfalls konstruiert und umfasst Pflege und Abstimmung außerhalb der einzelnen Anfragen. Von den 85 Minuten Unterschied bleiben unter dieser Annahme 55 Minuten mögliche Kapazität. Eine einmalige Einrichtung ist darin noch nicht enthalten.

Bei einem echten Pilot erfasst du zusätzlich externe Ausgaben mit ihrem Zeitraum und ihrer Nutzungsannahme. Halte Geld und Zeit zunächst in getrennten Einheiten. Eine spätere gemeinsame Bewertung braucht begründete Annahmen und darf Kosten nicht doppelt zählen. Dieses Booklet liefert keine Investitionsberatung und keine behauptete Amortisationsdauer.

Prüfe, wie empfindlich **das Ergebnis ist**

Eine kleine Änderung der Nacharbeit kann die Entscheidung stärker verändern als ein schnellerer erster Entwurf.

Korrekturvariante	65 statt 25 Minuten
KI-Weg dann	50 + 80 + 65 = 195 Minuten
Vorgangsdifferenz dann	240 − 195 = 45 Minuten
Nach Betreuung	45 − 30 = 15 Minuten

Steigen die Korrekturen in unserer Demo von 25 auf 65 Minuten, erhöht sich der KI-Weg auf 195 Minuten. Gegenüber 240 bleiben 45 Minuten. Nach denselben 30 Minuten Betreuung verbleiben nur noch 15 Minuten mögliche Nettokapazität. Die Rechnung ist vollständig konstruiert und stellt keine Prognose dar.

Mit dieser Variation erkennst du, welche Annahme dein Ergebnis trägt. Du brauchst keine Vielzahl fantasievoller Szenarien. Wähle die unsicheren Größen, die im Alltag plausibel schwanken könnten, und beobachte sie im nächsten begrenzten Versuch gezielt. Ein positiver Ausgangswert ist wenig hilfreich, wenn eine kleine unbekannte Nacharbeit ihn aufzehrt.



KI-gestaltetes persönliches Motiv. Keine Aufnahme eines realen Kundenprojekts.

Entscheide mit **der verbleibenden Unsicherheit**

Unser Rechenbeispiel zeigt einen möglichen Zeitvorteil, aber noch keine Wirkung im echten Betrieb. Die passende Entscheidung muss diesen Unterschied offen tragen und den nächsten Lernschritt begrenzen.

Du untersuchst jetzt, was die Beobachtung erklärt, für welche Fälle sie gelten könnte und welcher zusätzliche Nachweis eine Ausweitung rechtfertigen würde. Das Abschlussblatt hält auch einen begründeten Stopp als möglichen Ausgang fest.

Eine Veränderung hat **mehrere mögliche Ursachen**

Die schnellere Bearbeitung kann mit dem Pilot zusammenhängen und zugleich durch andere Änderungen beeinflusst werden.

Andere Quelle	Bereinigte Ablage erleichtert beide Wege
Andere Person	Erfahrung verändert Bearbeitungszeit
Andere Fälle	Einfachere Fragen verschieben das Mittel
Offene Frage	Welcher Anteil hängt am neuen Ablauf?

Vielleicht nutzt das Team während des Versuchs eine frisch bereinigte Regelablage. Vielleicht bearbeitet eine besonders erfahrene Person die KI-Fälle. Vielleicht ist die Fallmischung leichter. Jede dieser Erklärungen würde die Übertragbarkeit der Zeitdifferenz verändern. Die Auswertung muss solche Unterschiede sichtbar machen, statt sie nachträglich dem Modell zuzuschreiben.

Für einen echten Wirksamkeitsnachweis braucht es einen Vergleich, der die Zurechnung angemessen untersucht. Die HM-Treasury-Leitlinien behandeln dafür unterschiedliche Evaluationsansätze. Unser Lehrfall bleibt bewusst bei einer Rechnung mit beschriebenen Grenzen. Daraus folgt die Entscheidung für einen begrenzten nächsten Nachweis, keine Behauptung einer kausal belegten Produktivität.

Betrachte die Fälle **mit hoher Nacharbeit**

Ein brauchbarer Durchschnitt kann verbergen, dass der Pilot nur für einen Teil der Aufgaben passt.

Gruppe A	Vollständige eindeutige Regelanfrage
Gruppe B	Unvollständige oder mehrdeutige Anfrage
Heute unbekannt	Zeit und Qualität je Gruppe
Nächster Beleg	Einzelfälle mit Fallart dokumentieren

In unserer nächsten Demofrage trennen wir eindeutige Regelanfragen von Fällen mit unvollständiger Angabe. Für die zweite Gruppe könnte die Kontrolle länger dauern oder eine Rückfrage nötig werden. Wir erfinden dafür keine neuen Erfolgsquoten. Die vorhandenen Summen erlauben diese Unterteilung nicht, weil keine tatsächlichen Einzelaufzeichnungen vorliegen.

Genau diese Lücke wird zum Messauftrag. Eine spätere echte Auswertung soll Zeit und Abweichungen je zulässiger Fallart festhalten. Wenn nur eine Gruppe profitiert, kann ein engerer Einsatz sinnvoller sein als eine allgemeine Einführung. Die Entscheidung folgt dann dem beobachteten Umfang des Nutzens.

Erfasse nur, was **die Entscheidung braucht**

Die Messkarte verbindet einen Fall mit Zeitart, Qualität und Bedingungen, ohne unnötige persönliche Daten zu sammeln.

Fallkennung	Neutrale interne Kennung
Bedingungen	Weg, Stand und Fallart
Zeit	Aktive Minuten und getrennte Wartezeit
Qualität	Beobachtung mit prüfbarer Fundstelle
Lücke	Unbekannt, ausgeschlossen oder abgebrochen

Eine neutrale Fallkennung genügt für unsere Auswertung. Der Text der ursprünglichen Anfrage muss nicht in eine öffentlich geteilte Tabelle übernommen werden. Speichere notwendige Betriebsdaten nur im vorgesehenen geschützten Umfeld und kläre den zulässigen Zweck vor einer echten Erfassung. Dieses Booklet veröffentlicht ausschließlich Demo-Kennungen.

Trage Ausschlüsse und Unterbrechungen sichtbar ein. Wenn die Zeitmessung abbricht, bleibt die Zeit unbekannt und wird nicht als null gerechnet. Eine fehlende Qualitätsprüfung ist ebenso wenig eine bestandene Antwort. Diese einfachen Regeln verhindern, dass Lücken einen scheinbar besseren Durchschnitt erzeugen.

Unser **ausgefülltes Pilotprotokoll**

Der Lehrfall begründet einen begrenzten Messversuch und noch keine breite Einführung.

Rechenbefund	85 Minuten vor, 55 nach Demo-Betreuung
Unbekannt	Qualitätsverteilung und echte Einzelzeiten
Nicht behauptet	Kausaler Nutzen oder realisierte Geldersparnis
Entscheidung	Begrenzten Messversuch planen
Noch nicht	Breite Einführung freigeben

Die Demo zeigt rechnerisch 85 Minuten Unterschied bei zwanzig Anfragen und 55 Minuten mögliche Nettokapazität nach angenommener Betreuung. Zugleich fehlen echte Einzelzeiten, eine belastbare Qualitätsauswertung und ein Vergleich zur Ursachenklärung. Aus dieser Lage folgt, dass die Ausweitung noch nicht begründet ist.

Der ausgefüllte Download pilotplan-demo.txt begrenzt Fallarten, Zeitraum, Zuteilung und Qualitätsgrenze. Er legt auch fest, wann das fiktive Team fortsetzt, eingrenzt oder beendet. Diese vorab gesetzten Schwellen sind eigene Lehrentscheidungen. Sie sind weder ein allgemeiner Maßstab noch bereits beobachtete Ergebnisse.

Dein Protokoll **für die Pilotentscheidung**

Diese eigene Arbeitshilfe beginnt bei der Entscheidung, die du nach dem Versuch treffen möchtest.

Entscheidung und Zeitpunkt
Erwartete Wirkung und vermuteter Mechanismus
Vergleich, Fallarten und Zuteilung
Qualitätsmaßstab und Abbruchbedingung
Vollständiger Aufwand mit Zeitart und Zeitraum
Beobachtung, Unsicherheit und nächster Schritt

Trage dann die erwartete Wirkung, den Vergleich und die Qualitätsgrenze ein. Wähle nur Messgrößen, die eine dieser Fragen beantworten. Ein Dashboard mit vielen Nutzungszahlen kann hilfreich sein, liefert aber ohne Bezug zu Aufwand und Ergebnis noch keinen Nutzenbeleg. Bewahre Annahmen und tatsächliche Beobachtungen getrennt auf.

Die editierbare Vorlage nutzen-pilot.txt enthält zusätzlich die Demo-Rechnung und eine leere Messkarte. Bitte eine andere Person, die Rechnung aus deinen Angaben nachzuvollziehen. Wenn dafür unbekannte Annahmen erraten werden müssen, ist die Dokumentation noch nicht ausreichend für die beabsichtigte Entscheidung.

Fortsetzen ist nur einer **von drei Ausgängen**

Ein Pilot soll eine Handlung begründen und darf auch einen engeren Einsatz oder ein Ende ergeben.

Fortsetzen	Vereinbarte Bedingungen ausreichend belegt
Eingrenzen	Nutzen auf erkennbare Fallgruppe begrenzt
Beenden	Nutzen oder notwendige Grenze fehlt
Unser Lehrfall	Erst einen echten Vergleich planen

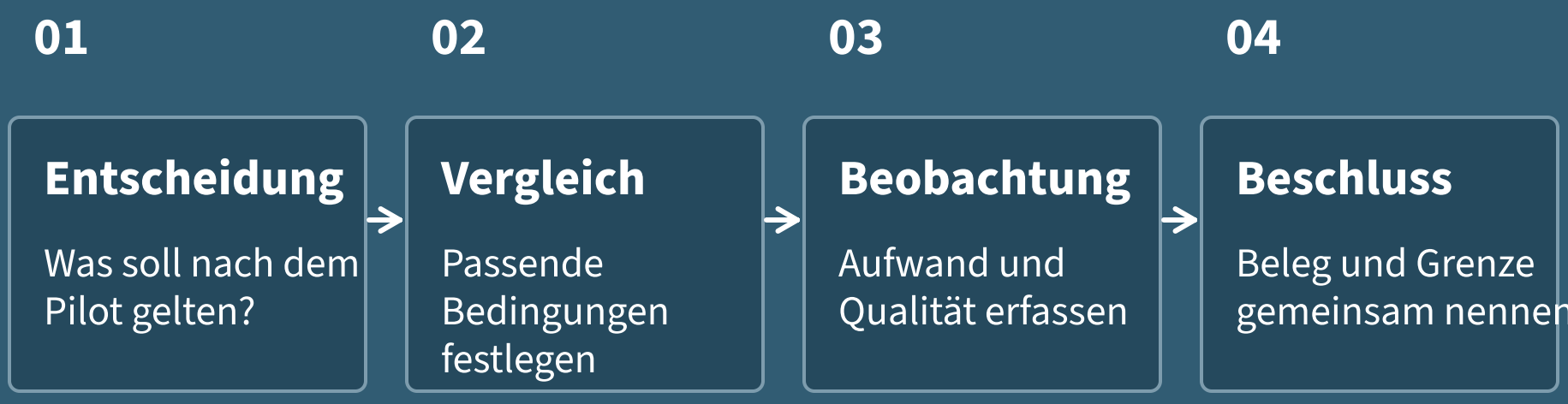
Fortsetzen passt, wenn die vorher vereinbarten Bedingungen ausreichend belegt sind. Eingrenzen kann passen, wenn nur bestimmte Fälle profitieren und diese zuverlässig erkennbar sind. Beenden ist sinnvoll, wenn der erwartete Nutzen ausbleibt oder notwendige Grenzen nicht eingehalten werden. Welche Belege ausreichen, entscheidet die zuständige Verantwortung vor dem Versuch.

In unserem Fall liegen noch gar keine echten Ergebnisse vor. Deshalb kann keine dieser Betriebsentscheidungen aus der Demo-Rechnung abgeleitet werden. Die einzige begründete nächste Handlung ist ein geeigneter Messplan oder der Verzicht auf einen Versuch, wenn seine Kosten und Folgen nicht zum möglichen Erkenntnisgewinn passen.

Frederik von der Heyden

So führt Messen **zu einer Entscheidung**

Die Messung wird vom geplanten Beschluss her aufgebaut und endet mit offen benannten Grenzen.



Unser Weg begann bei einer vermuteten Entlastung. Er zählte anschließend die vollständige Arbeit und zeigte den Einfluss zusätzlicher Betreuung. Danach blieb sichtbar, dass eine Rechnung aus Demo-Daten weder Qualität noch kausalen Nutzen belegt. Gerade diese Grenze macht den nächsten Schritt präziser: Es fehlen konkrete Beobachtungen, keine weitere Erfolgsgeschichte.

Übertrage diesen Ablauf auf eine kleine Aufgabe. Notiere die Entscheidung vor dem Start, bewahre die Fälle auf und betrachte die schwierigen Zeilen. Wenn du später eine Ausweitung begründest, soll eine andere Person sowohl den Nutzen als auch die verbleibende Unsicherheit aus denselben Belegen erkennen können.

Was unsere Rechnung **nicht zeigen kann**

Die Beispiele lehren Rechenschritte und Messfragen, sie berichten keine echte Verbesserung.

Nicht gemessen	Echter Zeitgewinn oder Qualitätsvorteil
Nicht nachgewiesen	Ursache einer beobachteten Veränderung
Nicht realisiert	Geldersparnis aus freien Minuten
Eigene Methodik	Demo-Rechnung und Messprotokoll

Zwanzig konstruierte Anfragen erlauben weder eine statistische Zuverlässigkeitsaussage noch eine Vorhersage für einen anderen Betrieb. Es wurden keine Unsicherheitsintervalle berechnet, weil kein real erhobener Datensatz vorliegt. Eine Prozentzahl aus diesen Summen wäre weiterhin nur eine Eigenschaft der Demo und kein vermarktbarer Produktivitätsbeleg.

Die finanzielle Entscheidung braucht tatsächliche Ausgaben, begründete Bewertungsannahmen und einen angemessenen Zeitraum. Bei erheblichen Investitionen oder Folgen solltest du die Planung mit qualifizierter methodischer und kaufmännischer Unterstützung prüfen. Die Arbeitshilfe kann diese Prüfung vorbereiten, sie ersetzt weder sie noch die Verantwortung für den realen Einsatz.

Quellen **für die Messprinzipien**

Die Quellen unterstützen den frühen Vergleich und passende Messgrößen, nicht das Ergebnis unseres Demofalls.

HM Treasury und Evaluation Task Force

Guidance on the Impact Evaluation of AI Interventions

www.gov.uk

HM Treasury und Evaluation Task Force

Magenta Book, Central Government guidance on evaluation

www.gov.uk

NIST

AI RMF Playbook, MEASURE 1.1

airc.nist.gov

Die KI-Evaluationsleitlinie von HM Treasury beschreibt unter anderem Wirkungsannahmen, Ausgangsmessung und die Auswahl eines angemessenen Evaluationsansatzes. Das Magenta Book trennt Fragen zu Umsetzung, Wirkung und Kosten-Nutzen-Betrachtung. NIST MEASURE 1.1 fordert passende Messansätze und die Dokumentation nicht gemessener Risiken.

Alle Quellen wurden am 17. November 2026 live geprüft. Unsere 240, 155 und 30 Minuten sind ausschließlich konstruierte Rechengrößen. Kein Herausgeber hat diese Zahlen erhoben oder die Vorlage als wirksam bestätigt. Die Quellenbelege trennen deshalb Grundprinzip, eigene Methodik und offene Fragen ausdrücklich. Die Fußnoten verweisen auf Grundprinzipien; Fälle, Formulare und Rechnungen sind eigene Methodik.

Miss **einen Arbeitsweg vollständig**

Beginne mit einer konkreten Entscheidung und der Frage, welcher Beleg sie verändern würde.

Wenn du Suche, Entwurf, Kontrolle und Nacharbeit gemeinsam betrachtest, wird ein Pilot leichter beurteilbar. Halte daneben die Qualitätsgrenze sichtbar und frage, wofür eine mögliche Entlastung tatsächlich genutzt werden kann. So erhält auch eine zurückhaltende Entscheidung eine nachvollziehbare Grundlage.

Dieses kostenlose Praxisbooklet von Frederik von der Heyden wurde KI-gestützt redigiert. Alle Zahlen sind neutrale Demo-Daten. Die persönlichen Motive sind KI-gestaltet und gekennzeichnet. Den Kontakt für ein Gespräch über einen konkreten Ablauf findest du auf der Website.

frederikvonderheyden.de