

KI-Bildgenerierung im Unternehmen – Die strategische Alternative zur Cloud



Die Entscheidung fällt nicht an der Grafikkarte, sondern am Flipchart. Dieses Beitragsbild ist KI-generiert und entsprechend gekennzeichnet.

WISSEN <u>KI-Bildgenerierung – der Überblick</u>	BERATUNG <u>Beratung zur KI-Bildgenerierung</u>	SCHULUNG <u>Schulung KI-Bildgenerierung</u>
Das gesamte Themenfeld von Diffusion bis Governance an einem Platz.	Betriebsmodell, Kostenrahmen und Governance für deinen konkreten Fall.	Damit die Leute, die das Budget verantworten, wissen, worüber sie entscheiden.

Die Frage kommt selten aus der IT. Sie kommt aus dem Marketing, meistens an einem Donnerstagnachmittag, und sie klingt harmlos: „Wir brauchen mal ein paar Abos für diese Bild-KI.“ Vier Wochen später liegen elf Abos auf elf verschiedenen Kreditkarten, niemand weiß, welche Bilder mit welchem Werkzeug entstanden sind, und die Rechtsabteilung fragt, ob die Produktskizzen für das nächste Geschäftsjahr eigentlich schon auf fremden Servern liegen. Sie liegen dort. Natürlich liegen sie dort.

Das ist keine Technikgeschichte. Das ist eine Beschaffungsgeschichte, die schiefging, weil sie nie als Beschaffungsentscheidung behandelt wurde. Genau darum geht es hier: nicht darum, welches Modell die schöneren Hände malt, sondern darum, welches Betriebsmodell du für die nächsten drei Jahre unterschreibst. Ob du Bildproduktion einkaufst oder selbst betreibst, ist dieselbe Art von Entscheidung wie die Frage, ob die Firma eine eigene Druckerei unterhält oder zur Druckerei fährt. Man kann sie so oder so beantworten. Man sollte sie nur nicht aus Versehen beantworten.

Der Satz, den du in Präsentationen dauernd hörst, lautet: lokal ist sicherer, Cloud ist billiger. Beide Hälften sind bestenfalls halbrichtig. Lokal ist nicht automatisch sicher, sondern erst dann, wenn jemand die Zugriffe protokolliert und die Modelle pflegt. Und Cloud ist nicht automatisch billiger,

sondern nur bis zu einem Punkt, den man ausrechnen kann. Diesen Punkt rechne ich weiter unten aus, mit offengelegten Annahmen, die du mir jederzeit streichen darfst.

Die technische Seite der Abwägung – Speicherbedarf, Modellfamilien, Geschwindigkeit – steht in [Cloud vs. lokale KI-Bildgenerierung](#). Wenn du das ganze Themenfeld von Grund auf brauchst, von der Funktionsweise bis zur Governance, findest du es im Überblick [KI-Bildgenerierung – der Überblick](#). Dieser Beitrag hier nimmt bewusst die Perspektive der Person ein, die am Ende unterschreibt: Was kostet es über drei Jahre, wovon macht es dich abhängig, was ist es wert, und in welcher Reihenfolge gehst du vor, ohne dich zu blamieren.

Die Entscheidung, die du wirklich triffst

„Cloud oder lokal“ ist eine schlechte Frage, weil sie zwei Dinge vermischt, die nichts miteinander zu tun haben: wo die Rechenleistung steht und wem der Prozess gehört. Du kannst eine Grafikkarte im eigenen Serverraum betreiben und trotzdem keinerlei Kontrolle über deine Bildproduktion haben, weil niemand dokumentiert, mit welchem Modell welches Motiv entstanden ist. Und du kannst über eine fremde Schnittstelle arbeiten und trotzdem einen sauber versionierten Prozess haben, weil du jeden Aufruf mit Modellkennung, Parametern und Freigabe protokollierst.

Die brauchbare Frage lautet deshalb: Welche Teile der Wertschöpfung willst du besitzen? Es gibt drei davon. Erstens die Rechenleistung – die Karte, die die Bilder erzeugt. Zweitens das Modell – die Gewichte, die entscheiden, wie die Bilder aussehen. Drittens den Prozess – die Reihenfolge aus Vorlage, Erzeugung, Prüfung, Freigabe, Ablage. Du kannst jeden dieser drei Teile einzeln kaufen oder selbst halten, und genau aus diesen Kombinationen entstehen die Wege, die in der Praxis vorkommen.

Drei Wege, nicht zwei

In fast jedem Projekt tauchen dieselben drei Varianten auf, und die mittlere wird regelmäßig übersehen, weil sie in keine der beiden Lagerdiskussionen passt.

- **Weg A – Abos und Schnittstellen.** Du kaufst Plätze bei einem oder mehreren Anbietern und rufst für Serien eine Programmierschnittstelle auf. Rechenleistung, Modell und ein Teil des Prozesses kommen von außen. Startzeit: ein Nachmittag.
- **Weg B – eigene Grafikkarte.** Die Karte steht im Haus, die Modellgewichte liegen auf deinem Speicher, der Prozess gehört dir vollständig. Startzeit: Wochen, und zwar Wochen, in denen jemand nichts anderes macht.
- **Weg C – gemietete GPU-Instanz.** Du mietest die Karte stunden- oder monatsweise, installierst darauf aber deine eigene Strecke mit deinen eigenen Modellgewichten. Das Blech ist fremd, Modell und Prozess sind deine. Startzeit: ein paar Tage.

Weg C ist der interessanteste, weil er die Diskussion entschärft. Er trennt die Frage „gehört uns der Prozess?“ sauber von der Frage „gehört uns das Blech?“. In Projekten, in denen die Beschaffung neun Monate für eine Hardwarebestellung braucht, ist das gelegentlich der einzige Weg, der innerhalb eines Geschäftsjahres irgendwo ankommt.

	Weg A: Abos und API	Weg C: GPU-Instanz	Weg B: eigene Karte
Rechenleistung	fremd	gemietet	eigen
Modellgewichte	fremd, nicht einsehbar	eigen, einfrierbar	eigen, einfrierbar
Prozess und Reihenfolge	teilweise fremd	eigen	eigen
Zeit bis zum ersten Bild	Stunden	Tage	Wochen
Kostenverlauf	läuft mit der Nutzung mit	überwiegend fix, Rest	fast vollständig fix

		nach Laufzeit	
Vertrauliches Material	geht nach draußen	bleibt beim Betreiber der Instanz	bleibt im Haus
Spitzenlast abfangen	trivial	trivial	teuer oder gar nicht
Wer haftet für den Betrieb	der Anbieter	geteilt	du

Was der Bestellvorgang verdeckt

Der Grund, warum Weg A in Unternehmen so oft ungeplant entsteht, ist banal: Ein Abo für dreißig oder sechzig Dollar im Monat liegt unterhalb jeder Genehmigungsschwelle, die es in einem mittelständischen Unternehmen gibt. Es wird also nicht entschieden, sondern gekauft. Elf Einzelkäufe ergeben am Ende einen Jahresbetrag, über den man diskutiert hätte, wenn er auf einer einzigen Rechnung gestanden hätte – und gleichzeitig eine Landschaft, die niemand überblickt.

Ich habe das oft genug gesehen, um es als Regel zu formulieren: Die erste Aufgabe in solchen Projekten ist nicht die Auswahl eines Werkzeugs, sondern eine Inventur. Wer nutzt was, wofür, mit welchen Daten, auf wessen Rechnung. Das dauert zwei Tage und verändert die Diskussion mehr als jede Marktübersicht. Welche Anbieter es derzeit überhaupt gibt und wie sie sich unterscheiden, steht in der [Übersicht der wichtigsten KI-Bildgeneratoren](#).

FAKTEN • Der Markt, Stand September 2026

Die Abo-Tarife der großen Bilddienste liegen zwischen rund zehn und rund hundertzwanzig US-Dollar je Platz und Monat, bei jährlicher Zahlung typischerweise zwanzig Prozent darunter. Die kommerzielle Nutzung der Ergebnisse ist bei den etablierten Anbietern in allen bezahlten Stufen enthalten – die Möglichkeit, nicht öffentlich zu generieren, dagegen erst in den oberen Stufen. Über Programmierschnittstellen wird nach Auflösung abgerechnet: rund 0,03 US-Dollar je Bild bei 1K, rund 0,05 bei 2K und rund 0,08 bei 4K sind die gängige Spanne; FLUX.2 [pro] rechnet mit etwa 0,03 US-Dollar je Megapixel ähnlich. Auf der Hardwareseite sieht es umgekehrt aus: Eine GeForce RTX 5090 startete Anfang 2025 mit einer Preisempfehlung von 2.099 Euro und ist im September 2026 wegen der Knappheit bei GDDR7-Speicher unterhalb von 5.200 Euro praktisch nicht zu bekommen. Die Profikarte RTX PRO 6000 Blackwell mit 96 GB wurde im August 2026 von rund 8.500 auf 16.000 US-Dollar Listenpreis angehoben.

Was drei Jahre wirklich kosten

Jetzt wird gerechnet. Und zwar über drei Jahre, weil das der Zeitraum ist, über den Hardware abgeschrieben wird und über den sich Abos zu einer Summe addieren, die auffällt. Ein Jahresvergleich macht die Cloud immer zum Sieger, ein Fünfjahresvergleich immer die eigene Karte. Drei Jahre sind ehrlich.

Die Annahmen, die du mir streichen darfst

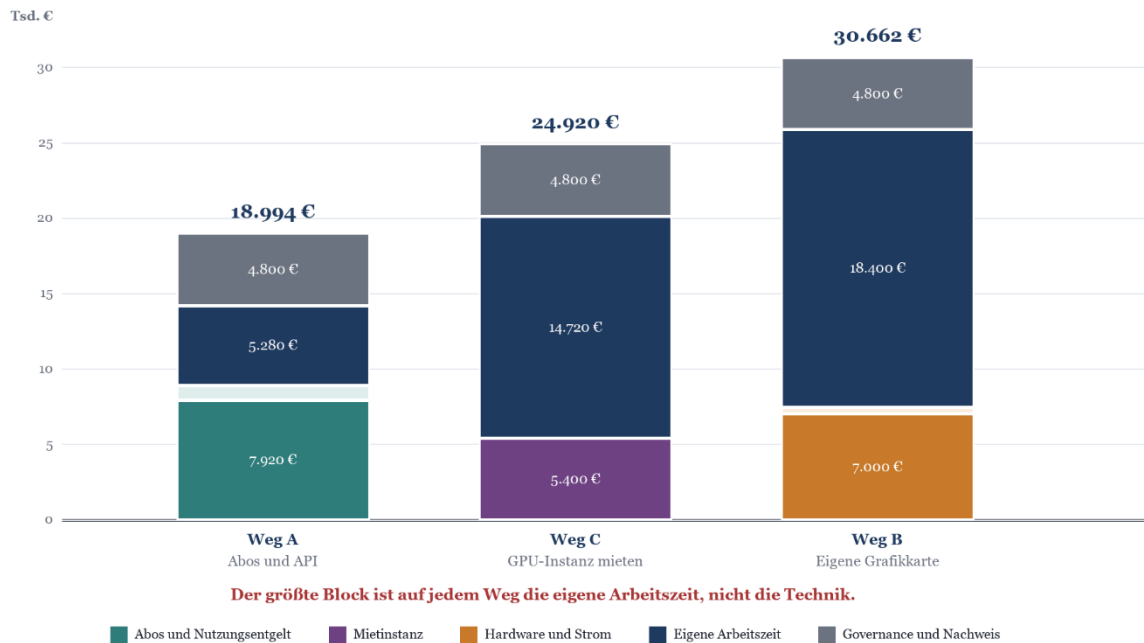
Jede Modellrechnung ist nur so gut wie ihre Annahmen, deshalb stehen sie hier vollständig und nicht in einer Fußnote. Angenommen ist ein Fachbereich mit fünf Plätzen, der 600 veröffentlichte Bilder pro Jahr braucht. Der Ausschussfaktor liegt bei zwölf: Für jedes Motiv, das am Ende verwendet wird, entstehen zwölf Generierungen aus Varianten, Verwürfen und Korrekturläufen. Das ergibt 7.200 Generierungen pro Jahr und 21.600 über drei Jahre. Interne Arbeitszeit ist mit 800 Euro je Personentag angesetzt, Gewerbestrom mit 25 Cent je Kilowattstunde, der Dollarkurs mit 0,92 Euro.

Zwei dieser Annahmen sind erfahrungsgemäß strittig. Der Ausschussfaktor zwölf klingt hoch, bis man ihn zum ersten Mal misst; in Teams ohne eingespielte Vorlagen habe ich auch schon zwanzig gesehen. Und die 800 Euro je Personentag sind ein Vollkostensatz, kein Gehalt. Wer stattdessen mit null rechnet, weil „das machen wir ja nebenbei“, bekommt eine hübschere Zahl und eine falsche Entscheidung.

Posten über 36 Monate	Weg A: Abos und API	Weg C: GPU-Instanz	Weg B: eigene Karte
Hardware oder Miete	—	5.400 € (2.640 Stunden zu rund 1,84 €, plus Speicher)	7.000 € (Karte rund 5.250 €, Restsystem 1.750 €)
Abos, 5 Plätze	7.920 € (5 × 44 € × 36 Monate)	—	—
Generierung nach Verbrauch	994 € (21.600 Bilder zu rund 0,046 €)	—	—
Strom	—	im Mietpreis enthalten	462 € (0,7 kW, 4 h je Arbeitstag)
Aufbau und Betrieb	5.280 € (6,6 Personentage)	14.720 € (18,4 Personentage)	18.400 € (23 Personentage)
Governance und Nachweis	4.800 € (6 Personentage)	4.800 €	4.800 €
bSumme	brund 19.000 €	brund 24.900 €	brund 30.700 €

Wohin drei Jahre Budget tatsächlich fließen

Modellrechnung für fünf Plätze und 7.200 Generierungen pro Jahr, Annahmen im Fließtext



Skizze 1: Die drei Wege nach Kostenarten aufgeschlüsselt. Auffällig ist nicht die Reihenfolge der Summen, sondern die Farbverteilung.

Das erste Ergebnis ist unbequem für beide Lager. Bei dieser Menge gewinnt der Abo-Weg deutlich, und zwar nicht knapp, sondern um mehr als ein Drittel. Das zweite Ergebnis ist wichtiger: Auf jedem der drei Wege ist der größte Einzelelement die eigene Arbeitszeit. Beim Eigenbetrieb macht sie rund sechzig Prozent der Dreijahressumme aus. Die Grafikkarte, über die in der Vorlage drei

Wochen diskutiert wird, ist rund ein Viertel. Der Strom, über den in der Vorlage ein ganzer Absatz steht, ist knapp anderthalb Prozent. Wer über Bildgenerierung im Unternehmen redet, redet also in Wahrheit über Personalkapazität und behauptet, über Technik zu reden.

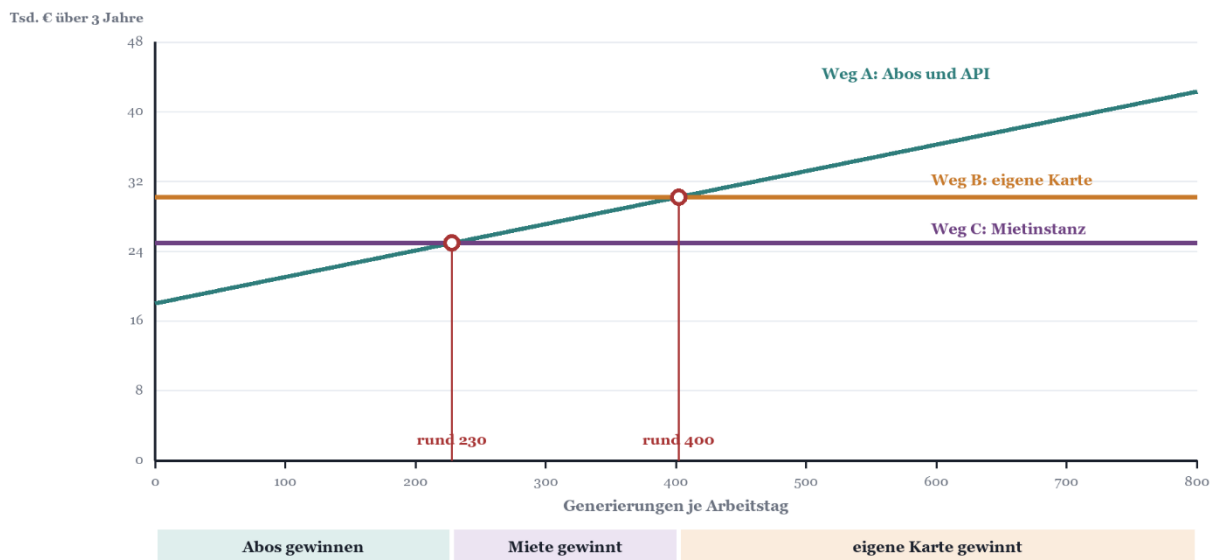
Und noch etwas fällt auf: Die reine Rechenleistung für 21.600 Bilder kostet über die Schnittstelle knapp tausend Euro. In drei Jahren. Das ist weniger, als ein einziger Workshop kostet. Die Vorstellung, man spare durch den Eigenbetrieb an den Generierungskosten, ist bei dieser Menge schlicht falsch – man spart an den Plätzen, nicht an den Bildern.

Der Punkt, an dem es kippt

Interessant wird es, wenn man die Menge variabel macht. Die Abo-Kosten für fünf Plätze bleiben konstant, die Generierungskosten wachsen linear mit der Menge, und die Kosten der eigenen oder gemieteten Karte bleiben nahezu unverändert, weil eine ausgelastete Karte fast nichts kostet außer Strom. Irgendwo kreuzen sich die Linien.

Ab wann sich die eigene Maschine rechnet

Dreijahreskosten gegen Generierungen je Arbeitstag, bei gleichbleibend fünf Plätzen



Verschiebt sich der Bildpreis von 0,03 auf 0,08 US-Dollar, wandert die rechte Marke von rund 660 auf rund 250.

Wer links der linken Marke arbeitet, kauft mit der eigenen Karte keine Ersparnis, sondern Kontrolle.

Skizze 2: Dieselbe Modellrechnung, aufgetragen über die Menge. Die beiden roten Marken sind die Punkte, an denen die Reihenfolge wechselt.

Bei etwa 230 Generierungen je Arbeitstag zieht die gemietete Instanz am Abo-Weg vorbei, bei etwa 400 die eigene Karte. Das sind keine astronomischen Zahlen: 400 Generierungen am Tag entsprechen bei einem Ausschussfaktor von zwölf rund 33 veröffentlichten Bildern täglich. Eine Agentur mit mehreren Kunden erreicht das ohne Anstrengung, eine Unternehmenskommunikation mit zwei Leuten nie. Deshalb ist die ehrliche Antwort auf „lohnt sich eine eigene Maschine?“ nicht ja oder nein, sondern eine Rückfrage nach der Menge.

Die Lage ist zudem empfindlich gegenüber dem Bildpreis. Arbeitest du konsequent in 1K und zahlst 0,03 US-Dollar je Bild, wandert die rechte Marke auf rund 660 Generierungen je Arbeitstag – die eigene Karte lohnt dann fast nie. Brauchst du 4K und zahlst 0,08, rutscht sie auf rund 250. Wer also die Modellrechnung für sein Haus aufstellt, muss zuerst wissen, in welcher Auflösung tatsächlich produziert wird. Das ist übrigens die Stelle, an der in der Praxis am meisten Geld verbrennt: Teams generieren gewohnheitsmäßig in der höchsten Stufe und skalieren anschließend auf Webgröße herunter.

WARNUNG · Du kaufst Hardware gerade im Hoch

Die Preise für Grafikkarten mit viel Speicher sind 2026 nicht gestiegen, sie sind explodiert. Eine Karte, die vor anderthalb Jahren gut zweitausend Euro kostete, liegt inzwischen jenseits von fünftausend; beim Profisegment hat der Hersteller den Listenpreis innerhalb eines Jahres fast verdoppelt. Ursache ist die Knappheit bei schnellem Grafikspeicher, nicht ein Sprung in der Leistung. Das heißt für dich: Eine Beschaffungsentscheidung, die vor zwei Jahren eindeutig für den Eigenbau ausgefallen wäre, fällt heute anders aus – und kann in achtzehn Monaten wieder anders ausfallen. Wer heute unbedingt eigene Rechenleistung braucht, sollte deshalb ernsthaft prüfen, ob die gemietete Instanz die Brücke ist, bis der Speichermarkt sich beruhigt hat.

TIPP · Miss den Ausschussfaktor, bevor du irgendetwas kaufst

Die eine Zahl, die deine gesamte Rechnung trägt, ist der Ausschussfaktor: wie viele Generierungen es braucht, bis ein Bild veröffentlicht wird. Sie steht in keiner Marktstudie, weil sie bei euch anders ist als überall sonst. Sie zu erheben kostet zwei Wochen und ein einfaches Zählprotokoll. Alles, was du danach rechnest, ist belastbar; alles, was du vorher rechnest, ist Prosa. Meine eigenen Zahlen stammen aus der eigenen Produktion – warum ich das so halte, steht in [Dogfooding: Ich setze ein, was ich berate](#).

Wovon du abhängig bist – und wovon nicht

„Anbieterbindung“ ist eines dieser Wörter, die in Vorlagen gut aussehen und in Diskussionen nichts erklären. Denn gebunden bist du immer, die Frage ist nur, woran. Es hilft, drei Arten von Bindung auseinanderzuhalten, weil sie unterschiedlich teuer und unterschiedlich schnell wirksam werden.

Drei Arten von Bindung

Preisbindung ist die harmloseste. Ein Anbieter erhöht die Tarife, du zahlst mehr oder gehst. Das tut weh, ist aber kalkulierbar und du merkst es sofort, weil es auf der Rechnung steht.

Prozessbindung ist teurer. Sie entsteht nicht durch Verträge, sondern durch Gewohnheit: Vorlagen, Textbausteine, Referenzbilder, abgestimmte Einstellungen, eine eingespielte Arbeitsweise. Nichts davon lässt sich exportieren. Wenn du den Anbieter wechselst, wechselst du nicht das Werkzeug, sondern die Handschrift deines Teams – und die Einarbeitung dauert länger als die Migration.

Ergebnisbindung ist die gefährlichste, weil sie lautlos wirkt. Dein Bildkatalog ist in einem bestimmten Stil entstanden. Wird das Modell dahinter geändert oder abgekündigt, kannst du zwar weiterhin Bilder erzeugen, aber keine, die zu den bereits veröffentlichten passen. Das merkst du erst, wenn ein Bestandsmotiv nachgeliefert werden soll – typischerweise zwei Tage vor Drucktermin.

Bindungsart	Weg A: Abos und API	Weg C und B: eigene Strecke
Preis	Anbieter setzt Tarife, Anpassung jederzeit möglich	Hardware- oder Mietpreis bei Beschaffung fixiert
Prozess	Vorlagen und Einstellungen bleiben beim Anbieter	Vorlagen liegen als Dateien bei dir, versionierbar
Ergebnis	Modellstand ändert sich ohne dein Zutun	Modellversion einfrierbar, solange du sie aufbewahrst
Verfügbarkeit	Abkündigung möglich, Fristen setzt der	läuft weiter, solange die Karte läuft

	Anbieter	
Recht am Modell	Nutzungsbedingungen des Anbieters, änderbar	Modelllizenzen, gilt für die heruntergeladene Version fort
Rechtsschutz bei Ansprüchen Dritter	bei einzelnen Anbietern vertraglich zugesagt, mit Deckelung und Auflagen	trägst du vollständig selbst

Die letzte Zeile wird regelmäßig unterschätzt und ist das stärkste Argument, das für den Einkauf spricht. Einzelne Anbieter – Adobe ist hier am weitesten – sagen ihren zahlenden Kunden eine Freistellung gegen Urheberrechtsansprüche Dritter zu, im Unternehmenstarif mit höheren Deckelungen. Diese Zusage ist kein Freibrief: Sie erlischt, wenn gegen die Nutzungsrichtlinien verstoßen wird, etwa indem reale Personen, Marken oder geschützte Figuren im Auftrag genannt werden, und sie deckt in der Regel Urheberrecht ab, nicht Marken- und Persönlichkeitsrecht. Trotzdem: Wer selbst generiert, hat diese Zusage nicht. Er trägt das Risiko allein. Das ist eine Position, die in keiner Kostentabelle steht und die du bewusst übernehmen solltest, nicht versehentlich.

Auch offene Modelle haben Eigentümer

Der zweite Irrtum betrifft die andere Seite. „Offene Gewichte“ heißt nicht „herrenlos“. Die Lizenzlage bei lokal betreibbaren Modellen ist im Jahr 2026 differenzierter, als es die Diskussion vermuten lässt, und sie entscheidet darüber, ob du die erzeugten Bilder überhaupt verkaufen darfst.

WICHTIG · Lizenzschwellen, die im Beschaffungsprozess auftauchen sollten

Bei Black Forest Labs liegt FLUX.2 [klein] in der 4-Milliarden-Variante unter Apache 2.0 und ist damit kommerziell frei nutzbar; die 9-Milliarden-Variante und FLUX.2 [dev] mit 32 Milliarden Parametern stehen unter der nicht kommerziellen FLUX-Lizenz und brauchen für den geschäftlichen Einsatz eine gesonderte Vereinbarung. Die Community-Lizenz von Stability AI erlaubt die kommerzielle Nutzung frei, solange die Organisation unter einer Million US-Dollar Jahresumsatz liegt – darüber wird eine Unternehmenslizenz fällig, unabhängig davon, wie wenig Bilder erzeugt werden. Ein Umsatzsprung kann dich also lizenzpflichtig machen, ohne dass sich an deinem Betrieb irgendetwas geändert hat. Welche Modelle derzeit unter welcher Lizenz stehen, steht in [Die 15 wichtigsten Modelle zur lokalen Bildgenerierung](#).

Daraus folgt eine unbequeme Erkenntnis für alle, die den Eigenbetrieb mit Unabhängigkeit begründen: Du tauschst die Abhängigkeit von einem Dienst gegen die Abhängigkeit von einer Lizenz. Der Unterschied ist allerdings real und spricht für den Eigenbetrieb: Eine Lizenz, unter der du eine Datei heruntergeladen hast, gilt für diese Version fort. Ein Dienst kann seine Bedingungen mit Wirkung für morgen ändern. Die heruntergeladene Modelldatei von heute funktioniert auch dann noch, wenn der Hersteller seine Meinung ändert – vorausgesetzt, du hast sie aufgehoben und weißt, welche es war. Was genau in so einer Datei steckt, steht in [Checkpoints, Modelle und VAE](#).

Abkündigung ist der unterschätzte Fall

Der Fall, an den bei Vertragsabschluss niemand denkt, ist nicht die Preiserhöhung, sondern das Verschwinden. Modelle werden abgekündigt, und zwar mit Fristen, die nach Entwicklerlogik großzügig und nach Unternehmenslogik knapp sind. Das erste Bildmodell von OpenAI, im Frühjahr 2025 noch der Stand der Technik, wird am 23. Oktober 2026 abgeschaltet. Wer seinen Markenlook darauf aufgebaut hat, hat keine achtzehn Monate Übergangszeit, sondern eine Nachricht in einem Entwicklerblog.

WARNUNG · Die Nachlieferung ist der Stresstest, nicht die Erstproduktion

Frage jeden Anbieter und jedes interne Vorhaben dasselbe: Kann ich in zwei Jahren ein Motiv nachliefern, das zu einem heute veröffentlichten passt? Bei einem Dienst lautet die ehrliche Antwort fast immer: vermutlich nicht, weil das Modell bis dahin mehrfach aktualisiert wurde. Bei eigener Strecke lautet sie: ja, wenn du Modellversion, Vorgabe und Startwert aufbewahrst. Das ist der praktische Kern der Reproduzierbarkeit, und er hängt an unspektakulären Dingen – nachzulesen in [Der Seed bei der KI-Bildgenerierung](#) und [batch size und Reproduzierbarkeit](#). Wer beides nicht protokolliert, hat auch lokal keine Reproduzierbarkeit, sondern nur das gute Gefühl, sie zu haben.

Markenkonsistenz und Governance: die beiden Werte, die in keiner Rechnung stehen

Wenn der Kostenvergleich bei mittleren Mengen so klar für den Einkauf ausfällt – warum betreiben dann überhaupt Unternehmen eigene Bildstrecken? Weil zwei Positionen in der Tabelle fehlen, die sich schlecht beziffern lassen und trotzdem Geld sind.

Markenkonsistenz ist keine Geschmacksfrage, sondern eine Kostenposition

Ein Bild, das nicht zur Marke passt, ist kein leicht danebenliegendes Bild. Es ist Ausschuss. Es kostet die Generierung, die Sichtung, die Diskussion in der Abstimmung und am Ende die Neuproduktion. Der teuerste Teil davon ist die Sichtung, weil sie Menschen bindet, die an dem Tag etwas anderes hätten tun sollen.

Rechne es an der Modellrechnung von oben durch. Bei 600 veröffentlichten Bildern im Jahr und einem Ausschussfaktor von zwölf werden 7.200 Generierungen gesichtet. Senkt ein eingelernter Hausstil den Faktor von zwölf auf fünf, entfallen 4.200 Sichtungen im Jahr. Bei zwanzig Sekunden je Sichtung sind das gut 23 Stunden, also knapp drei Personentage, also rund 2.300 Euro pro Jahr und rund 7.000 Euro über drei Jahre. Das ist ungefähr der Preis der Grafikkarte, über die vorne so lange diskutiert wurde – und es ist eingespart, ohne dass ein einziges Bild günstiger geworden wäre.

Der Punkt ist nicht, dass Markenkonsistenz nur lokal möglich wäre. Auch Dienste kennen Referenzbilder, Stilvorgaben und wiederverwendbare Beschreibungen. Der Unterschied ist, dass du diese Konsistenz dort nicht besitzt: Du kannst sie nicht einfrieren, nicht versionieren und nicht vererben, wenn jemand das Unternehmen verlässt. Auf eigener Strecke wird aus dem Hausstil eine Datei – ein eingelerntes Zusatzmodell, wie es in [Was ist ein LoRA?](#) beschrieben ist –, und Dateien lassen sich sichern, prüfen und weitergeben. Wie das konkret aussieht und wo die Grenzen liegen, zeige ich an meinem eigenen Material in [Markenkonsistente KI-generierte Bilder](#).

Wo der Bedarf tatsächlich entsteht

Bevor du irgendetwas entscheidest, solltest du wissen, wo die Bilder eigentlich gebraucht werden. In der Praxis sind es fast immer dieselben fünf Stellen, und sie unterscheiden sich stark darin, wie vertraulich das Material ist und wie viele Varianten gebraucht werden.

Bedarfsstelle	Menge	Vertraulichkeit	Was daraus folgt
Kampagnen und Werbemittel	hoch, in Serien	hoch bis zur Veröffentlichung	der klassische Fall für eine eigene oder gemietete Strecke
Unternehmenskommunikation, Beiträge, Rundschreiben	mittel, dauerhaft	gering	der klassische Fall für Abos

Arbeitgebermarke und Karriereseiten	niedrig, schubweise	mittel, oft Personenbezug im Auftrag	Auftragsinhalte prüfen, Personenbezug vermeiden
Schulungs- und Handbuchmaterial	mittel, wiederkehrend	hoch, weil interne Abläufe beschrieben werden	Auftragstexte verraten mehr als die Bilder
Agenturen und Dienstleister	sehr hoch, viele Varianten	sehr hoch, Verschwiegenheitsvereinbarungen	hier kippt die Rechnung regelmäßig zugunsten des Eigenbetriebs

Die vierte Zeile ist die, die in Projekten am häufigsten übersehen wird. Nicht das erzeugte Bild ist vertraulich, sondern der Auftragstext, mit dem es bestellt wurde. Wer eine Schulungsgrafik für einen internen Freigabeprozess bestellt, beschreibt in der Bestellung den internen Freigabeprozess. Diese Texte sind oft die heikelsten Daten im ganzen Vorgang, und sie werden mit einer Selbstverständlichkeit nach draußen geschickt, die einen staunen lässt.

Governance kostet auf beiden Wegen, nur unterschiedlich

Die verbreitete Vorstellung, der Einkauf von Diensten erspare einem die Governance, ist falsch. Er verschiebt sie. Seit dem 2. August 2026 gelten die Transparenzpflichten aus Artikel 50 der KI-Verordnung. Anbieter generativer Systeme müssen synthetische Inhalte maschinenlesbar kennzeichnen, etwa über Herkunftsdaten nach dem C2PA-Standard oder über Wasserzeichen. Betreiber – und das bist in aller Regel du – müssen offenlegen, wenn sie Deepfakes erzeugen oder veröffentlichte Texte von öffentlichem Interesse maschinell erstellen lassen. Die Kommission hat dazu einen freiwilligen Verhaltenskodex vorgelegt. Was das im Einzelnen bedeutet, steht in [EU AI Act: Was seit dem 2. August 2026 gilt](#); wie man KI-Bilder überhaupt erkennt, in [KI-generierte Bilder erkennen](#).

Pflicht oder Aufgabe	Weg A: Abos und API	Weg C und B: eigene Strecke
Maschinenlesbare Kennzeichnung der Ausgabe	liegt beim Anbieter; du musst prüfen, ob und wie sie erfolgt	liegt bei dir; die Kennzeichnung muss fest im Ablauf stecken
Offenlegung gegenüber dem Publikum	deine Aufgabe	deine Aufgabe
Auftragsverarbeitung und Drittlandtransfer	je Anbieter zu vereinbaren und jährlich nachzuhalten	entfällt für die Erzeugung selbst
Nachweis, womit ein Bild entstanden ist	nur so weit, wie der Anbieter protokolliert	vollständig, wenn du protokollierst
Lizenznachweis für Modelle	Nutzungsbedingungen, änderbar	eigener Modellkatalog mit Versionsstand
Löschkonzept und Aufbewahrung	abhängig von den Möglichkeiten des Anbieters	vollständig in eigener Hand, aber auch vollständig deine Arbeit

In der Modellrechnung stehen für Governance auf allen drei Wegen dieselben sechs Personentage. Das ist bewusst so gesetzt: Der Aufwand verschwindet nicht, er wandert nur. Auf dem Abo-Weg steckt er in der Anbieterprüfung, den Verträgen und der jährlichen Nachkontrolle. Auf der eigenen Strecke steckt er im Modellkatalog, in der Protokollierung und in der Kennzeichnung. Wer glaubt, einen der beiden Posten einsparen zu können, spart ihn nicht ein, sondern verschiebt ihn in die Zukunft, wo er teurer ist.

TIPP · Der Nachweis braucht einen Ort, nicht einen Ordner

Ob Modellkatalog, Lizenzstand oder Kennzeichnungsnachweis – das alles wird nur geführt, wenn es einen festen Platz hat, an dem Änderungen sichtbar werden. In meinen Projekten läuft das über UB.Dossier: Werte und Versionsstände werden dort gepflegt, der Katalog bleibt an einer Stelle, und der Nachweis fällt als Nebenprodukt der Pflege an statt als Sonderaktion vor der Prüfung. Wer lieber in der vorhandenen Umgebung bleibt, findet einen gangbaren Weg in [Digital Asset Management für KI-Bilder mit SharePoint Online und Power Platform](#).

Make or Buy – und wie du dort hinkommst

Aus allem bisherigen lässt sich die Entscheidung auf zwei Achsen eindampfen. Die eine ist die Menge, weil sie die Kosten bestimmt. Die andere ist die Vertraulichkeit des Materials, weil sie bestimmt, ob die günstigere Variante überhaupt zur Auswahl steht. Alles andere – Modellfamilie, Oberfläche, Betriebssystem – ist Ausführungsdetail und gehört nicht in eine Vorstandsvorlage.

Make or Buy: die zwei Fragen, die wirklich entscheiden

Alles andere ist Ausführungsdetail



Die Waagerechte entscheidet der Preis. Die Senkrechte entscheidet die Rechtsabteilung.

Skizze 3: Das Entscheidungsraster. Rechts unten wird gemischt gefahren, und das ist häufiger der richtige Platz, als es Grundsatzdiskussionen nahelegen.

Das Raster in Worten

Unten links, bei wenig Bildern und unkritischem Material, kaufst du Plätze ein und hörst auf, darüber nachzudenken. Jede Eigenbau-Diskussion kostet hier mehr Besprechungszeit, als sie je einspart. Oben links, bei wenig Bildern, aber vertraulichem Material, ist die eigene oder gemietete Strecke richtig – nicht, weil sie billiger wäre, sie ist es nachweislich nicht, sondern weil die Alternative ausscheidet. Rechts oben, viel Volumen und hohe Vertraulichkeit, ist der einzige Fall, in dem der Eigenbetrieb in beiden Dimensionen gewinnt; hier arbeiten Agenturen und Unternehmen mit Produktentwicklung unter Verschluss.

Rechts unten steht die Variante, die in Projekten am häufigsten das Richtige ist und am seltensten beschlossen wird: gemischt fahren. Ideensuche, Einzelmotive und Eiliges über eingekaufte Plätze; Serien, Markenmotive und alles, was in den Katalog wandert, über die eigene Strecke. Das klingt nach Kompromiss und ist in Wahrheit Arbeitsteilung. Der einzige Fallstrick: Es braucht eine klare Regel, welches Material auf welchem Weg bearbeitet werden darf, sonst hast du beide Kostenblöcke und keinen der Vorteile.

Frage an das eigene Haus	Antwort spricht für Einkauf	Antwort spricht für Eigenbetrieb
Wie viele Generierungen fallen je Arbeitstag an?	unter etwa 230	über etwa 400
Wie vertraulich sind Auftragstexte und Vorlagen?	unkritisch, nichts Unveröffentlichtes	Produktdaten, interne Abläufe, Kundenmaterial
Muss ein Motiv in zwei Jahren nachgeliefert werden?	nein, alles ist tagesaktuell	ja, Kampagnen laufen über Jahre
Gibt es eine Person mit Betriebsverantwortung?	nein, niemand hat Kapazität	ja, benannt und mit Zeitbudget
Wie hoch ist die Risikobereitschaft bei Rechtsansprüchen?	gering, Freistellungszusage gewünscht	tragbar, wird bewusst übernommen
Wie schnell muss es losgehen?	diesen Monat	dieses Halbjahr reicht
Wie stabil ist der Bedarf?	schwankt stark, Projektgeschäft	planbar und dauerhaft

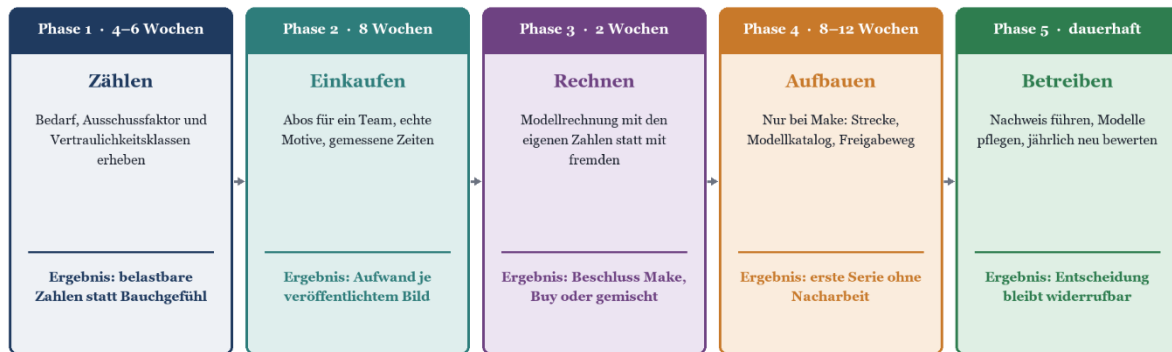
Die vierte Zeile ist ein Ausschlusskriterium, kein Punkt auf einer Liste. Wenn niemand benannt ist, der den Betrieb verantwortet, ist der Eigenbetrieb keine Option, sondern ein Projektrisiko mit Ankündigung. Ich habe mehr als eine eigene Bildstrecke gesehen, die nach acht Monaten stillstand, weil die Person, die sie aufgebaut hatte, in ein anderes Vorhaben wechselte und niemand die Modellverwaltung übernahm. Die Hardware stand dann noch zwei Jahre im Rack und wurde als Abschreibung geführt.

Der Einführungsweg in fünf Phasen

Der häufigste Fehler in solchen Vorhaben ist die Reihenfolge: Erst wird Hardware beschafft, dann wird überlegt, wofür. Der umgekehrte Weg kostet weniger und liefert schneller, weil er die teure Entscheidung so lange aufschiebt, bis Zahlen vorliegen.

Einführung in fünf Phasen, mit Abbruchstelle nach jeder

Wer alles auf einmal beschließt, beschließt vor allem im Dunkeln



Die teuerste Variante ist die, bei der Phase 4 beginnt, bevor Phase 1 fertig war.

Phase 2 kostet ein paar hundert Euro und beantwortet mehr Fragen als jede Marktrecherche.

Skizze 4: Fünf Phasen mit einer Abbruchstelle nach jeder. Nach Phase 3 steht die Entscheidung auf Zahlen aus dem eigenen Haus.

Phase 1, vier bis sechs Wochen: zählen. Wer erzeugt heute wie viele Bilder, wofür, mit welchen Werkzeugen und auf wessen Rechnung? Dazu der Ausschussfaktor und eine grobe Einteilung des Materials nach Vertraulichkeit. Ergebnis sind Zahlen, mit denen man rechnen kann – und fast immer eine Überraschung darüber, wie viele Abos es schon gibt.

Phase 2, acht Wochen: einkaufen. Ein Team bekommt ordentliche Plätze bei einem Anbieter und produziert echte Motive für echte Zwecke. Das ist kein Alibi-Schritt, sondern die billigste Art, den eigenen Aufwand je veröffentlichtem Bild zu messen. Ein paar hundert Euro Tarifkosten beantworten hier Fragen, die keine Marktrecherche beantwortet.

Phase 3, zwei Wochen: rechnen. Jetzt wird die Modellrechnung aus diesem Beitrag mit den eigenen Zahlen neu aufgestellt. Meine Annahmen sind Platzhalter, deine sind die Entscheidungsgrundlage. Am Ende steht ein Beschluss: einkaufen, selbst betreiben oder gemischt fahren – mit Begründung, die auch in sechs Monaten noch trägt.

Phase 4, acht bis zwölf Wochen: aufbauen. Nur nötig, wenn der Beschluss auf Eigenbetrieb lautet. Hier entstehen die Produktionsstrecke, der Modellkatalog mit Versionsständen und der Freigabeweg. Wie eine solche Strecke praktisch aussieht, steht in [Aufbau einer skalierbaren Bild-Maschine mit ComfyUI und Flux](#). Bei mir läuft die Produktion über UB.Forge: Arbeitsabläufe werden dort geändert, produziert wird daraus, und ein Plausibilitäts-Check läuft mit, bevor ein Ergebnis weitergereicht wird. Das Werkzeug ist austauschbar, das Prinzip nicht – ohne feste Abläufe hast du keine Strecke, sondern fünf Leute mit Geschmack.

Phase 5, dauerhaft: betreiben. Nachweis führen, Modelle pflegen, einmal im Jahr die Entscheidung neu bewerten. Der letzte Punkt ist der wichtigste und wird am häufigsten vergessen. Preise für Hardware und für Bilder bewegen sich derzeit so stark, dass eine Entscheidung aus dem Vorjahr eine Annahme ist, keine Tatsache.

TIPP · Die Befähigung vorher einplanen, nicht hinterher nachreichen

In der Modellrechnung stecken 23 Personentage Aufbau und Betrieb für die eigene Karte. Die setzen voraus, dass jemand weiß, was er tut. Ist das nicht der Fall, verdoppelt sich diese Zahl geräuschlos, und zwar in einem Posten, den niemand kontrolliert, weil er nicht auf einer Rechnung steht. Ein paar Tage Einarbeitung vorab sind die günstigste Position im ganzen Vorhaben – die Inhalte dazu stehen

in [Schulung KI-Bildgenerierung](#). Wenn du den Entscheidungsweg lieber mit jemandem durchgehst, der die Zahlen schon öfter aufgestellt hat, führt der Weg über [Beratung zur KI-Bildgenerierung](#).

Fazit

Die lokale Bildstrecke ist eine strategische Alternative zur Cloud – aber nicht aus den Gründen, die üblicherweise dafür angeführt werden. Sie ist bei üblichen Unternehmensmengen nicht billiger, im Gegenteil: In der Modellrechnung über drei Jahre liegt der Einkauf mit rund 19.000 Euro klar vor der gemieteten Instanz mit rund 24.900 und der eigenen Karte mit rund 30.700 Euro. Erst jenseits von etwa 400 Generierungen je Arbeitstag dreht sich das, und der Punkt verschiebt sich mit jedem Cent, den ein Bild über die Schnittstelle kostet.

Was die eigene Strecke tatsächlich liefert, sind drei Dinge, die man nicht kaufen kann: vertrauliches Material bleibt vertraulich, auch der Auftragstext. Der Markenlook wird zu einer Datei, die man sichern, versionieren und weitergeben kann, statt zu einer Gewohnheit, die mit der Person geht. Und ein Motiv lässt sich in zwei Jahren nachliefern, weil die Modellversion noch da ist. Wenn eines dieser drei Dinge für dein Geschäft zählt, ist der Aufpreis begründet. Wenn keines zählt, ist er Liebhaberei.

Der größte Posten bleibt in jedem Fall derselbe, egal welchen Weg du wählst: die eigene Arbeitszeit. Auf der eigenen Karte sind es rund sechzig Prozent der Dreijahressumme. Wer diese Zahl in der Vorlage auf null setzt, weil das Team es ja nebenbei macht, hat keine Entscheidung getroffen, sondern eine Rechnung gefälscht – und wird in achtzehn Monaten erklären müssen, warum die Bildproduktion stillsteht, seit eine bestimmte Person die Abteilung gewechselt hat.

Praktisch heißt das: Fang mit dem billigsten Weg an, miss dabei, und entscheide dann. Einge kaufte Plätze für acht Wochen kosten weniger als die Besprechungen, in denen über Grafikkarten diskutiert wird, und sie liefern die einzigen Zahlen, die zählen – deine eigenen. Alles Weitere folgt daraus fast von selbst. Die Grundlagen zum Thema, falls du sie für die Vorlage brauchst, stehen in [Basiswissen KI-Bildgenerierung: Wie Diffusion funktioniert](#).

Häufige Fragen

Lohnt sich eine eigene Grafikkarte bei den aktuellen Preisen überhaupt noch?

Seltener als noch vor zwei Jahren, und das liegt nicht an der KI, sondern am Speichermarkt. Karten mit viel Videospeicher haben sich seit Anfang 2025 im Preis mehr als verdoppelt, ohne entsprechend schneller geworden zu sein. In der Modellrechnung kippt die Entscheidung erst jenseits von rund 400 Generierungen je Arbeitstag zugunsten eigener Hardware. Wenn du unterhalb dieser Menge liegst und trotzdem eigene Rechenleistung brauchst – etwa weil das Material vertraulich ist –, ist die gemietete Instanz derzeit meist die vernünftige Brücke.

Was kostet der Einstieg mindestens?

Wenn du nur wissen willst, ob das Thema für euch trägt: ein paar hundert Euro und acht Wochen. Fünf ordentliche Plätze bei einem Anbieter kosten in dieser Zeit etwa die Hälfte eines Beratungstages. Dafür bekommst du echte Motive, gemessene Zeiten und einen Ausschussfaktor aus dem eigenen Haus. Jede Beschaffungsentscheidung, die vor diesen acht Wochen getroffen wird, beruht auf fremden Zahlen.

Können lokale Modelle qualitativ mit den großen Diensten mithalten?

Für die üblichen Geschäftszwecke ja. Der Abstand, der 2023 noch deutlich war, ist bei Werbemotiven, Illustrationen und Produktdarstellungen praktisch verschwunden. Wo Dienste weiterhin vorn liegen, ist das Verstehen sehr langer, verschachtelter Beschreibungen und die Textdarstellung im Bild. Umgekehrt liegen lokale Strecken vorn, wenn es um genaue Steuerung geht: Bildaufbau, Körperhaltung, Wiedererkennbarkeit über eine ganze Serie. Für die Entscheidung, um die es hier geht, ist Qualität ohnehin selten der Ausschlag – sie ist auf beiden Wegen ausreichend, und die Unterschiede liegen woanders.

Dürfen wir die erzeugten Bilder kommerziell nutzen?

Das hängt vom Weg ab und muss vor der ersten Veröffentlichung geklärt sein. Bei eingekauften Diensten ist die kommerzielle Nutzung in den bezahlten Tarifen üblicherweise enthalten, gelegentlich mit Auflagen zur Kennzeichnung. Bei lokal betriebenen Modellen entscheidet die Modelllizenz, und die ist uneinheitlich: Es gibt frei kommerzielle Modelle unter Apache 2.0, ausdrücklich nicht kommerzielle Varianten und Lizenzen mit einer Umsatzschwelle, oberhalb derer eine Unternehmenslizenz fällig wird. Ein Modellkatalog mit Lizenzspalte ist deshalb keine Bürokratie, sondern Voraussetzung.

Müssen wir KI-Bilder kennzeichnen?

Seit dem 2. August 2026 gelten die Transparenzpflichten der KI-Verordnung. Die maschinenlesbare Kennzeichnung synthetischer Inhalte trifft in erster Linie die Anbieter der Systeme; als Betreiber musst du offenlegen, wenn du Deepfakes veröffentlichst oder Texte von öffentlichem Interesse maschinell erstellen lässt. Betreibst du die Erzeugung selbst, rückst du näher an die Anbieterrolle heran – dann muss die Kennzeichnung fest in deinem Ablauf stecken und nicht am Ende händisch ergänzt werden. Unabhängig von der Rechtslage empfehle ich sichtbare Kennzeichnung: Sie kostet nichts und erspart Diskussionen.

Brauchen wir dafür ComfyUI oder ein vergleichbares Werkzeug?

Wenn du selbst betreibst: ja, in irgendeiner Form. Nicht weil das Werkzeug besonders schön wäre, sondern weil ein aus Bausteinen zusammengesetzter Ablauf als Datei existiert. Er lässt sich sichern, versionieren, weitergeben und einem Prüfer vorlegen. Genau das ist der Unterschied zwischen einer Strecke und fünf Leuten, die jeweils ihre eigenen Einstellungen im Kopf haben. Ob das Werkzeug am Ende so oder anders heißt, ist zweitrangig.

Können wir beides gleichzeitig betreiben?

Nicht nur können – in mittleren Häusern ist es meist die beste Variante. Ideensuche, Einzelmotive und Eiliges laufen über eingekaufte Plätze, Serien und Markenmotive über die eigene Strecke. Die einzige Bedingung ist eine klare, schriftliche Regel, welches Material auf welchem Weg bearbeitet werden darf. Ohne diese Regel hast du beide Kostenblöcke und keinen der beiden Vorteile.

Wie viele Leute brauchen wir dafür?

Für den Einkauf: eine Person, die Verträge und Nachweise führt, mit etwa sechs Personentagen über drei Jahre. Für den Eigenbetrieb: eine benannte Person mit einem halben Tag im Monat, plus fünf Tage Aufbau. Das klingt wenig und ist es auch – solange es jemand ist. Nicht benannte Verantwortung entspricht in der Praxis null Personentagen, und daran scheitern diese Vorhaben, nicht an der Technik.

Wie lange hält so eine Entscheidung?

Rechne mit zwölf bis achtzehn Monaten, dann gehört sie auf den Prüfstand. Die Hardwarepreise schwanken derzeit stärker als in den zwanzig Jahren davor, Bildpreise über Schnittstellen sinken tendenziell, Modelle werden mit kurzen Fristen abgekündigt. Deshalb gehört in jeden Beschluss

ein Satz dazu, unter welchen Bedingungen er neu bewertet wird. Ein Beschluss ohne Verfallsdatum ist in diesem Feld keine Strategie, sondern eine Wette.

WEITERLESEN

- › [KI-Bildgenerierung – der Überblick](#) – das gesamte Themenfeld im Zusammenhang
- › [Cloud vs. lokale KI-Bildgenerierung](#) – dieselbe Frage aus technischer Sicht
- › [Übersicht der wichtigsten KI-Bildgeneratoren](#) – wer derzeit am Markt ist und was er anbietet
- › [Die 15 wichtigsten Modelle zur lokalen Bildgenerierung](#) – Lizenzlage und Speicherbedarf der offenen Modelle
- › [Aufbau einer skalierbaren Bild-Maschine mit ComfyUI und Flux](#) – wie die eigene Produktionsstrecke praktisch aussieht
- › [Digital Asset Management für KI-Bilder mit SharePoint Online und Power Platform](#) – Katalog, Nachweis und Ablage an einem Ort
- › [Markenkonsistente KI-generierte Bilder](#) – wie aus Markenlook eine sicherbare Datei wird
- › [EU AI Act: Was seit dem 2. August 2026 gilt](#) – Transparenzpflichten und was sie für dich bedeuten
- › [Black Forest FLUX.2 in PowerPoint und Copilot](#) – wenn Bildgenerierung ohnehin in der Office-Umgebung ankommt
- › [Claude im Unternehmen](#) – dieselbe Make-or-Buy-Frage für Sprachmodelle
- › [Basiswissen KI-Bildgenerierung: Wie Diffusion funktioniert](#) – die Grundlagen für die Vorlage an das Gremium
- › [Dogfooding: Ich setze ein, was ich berate](#) – woher die Zahlen in diesem Beitrag stammen
- › [Beratung zur KI-Bildgenerierung](#) – Betriebsmodell und Kostenrahmen gemeinsam aufstellen
- › [Schulung KI-Bildgenerierung](#) – Befähigung, bevor der Betriebsaufwand aus dem Ruder läuft