# Ethisches Handeln durch Software Engineering

Thomas Matzner


*Ethische Fragen im Zusammenhang mit Informatiksystemen finden in den letzten Jahren zunehmendes Interesse. Ein Kernpunkt der von mir entworfenen Informatikethik (Matzner 2020) sind konkrete Moralempfehlungen, die denen, die an Entwicklung und Betrieb von Informatiksystemen beteiligt sind, zeigen, bei welchen Tätigkeiten welches Verhalten dazu dienen kann, moralisch erfreuliche Ergebnisse zu erzielen.*

An drei Fallstudien möchte ich einen Ausschnitt dieser moralischen Erwägungen illustrieren:

- (Nezik 2019) berichtet von einer Software, die von der österreichischen Arbeitsmarktverwaltung eingesetzt wird, um die Vermittlungschancen von Langzeitarbeitslosen zu berechnen. Um die Gelder für Ausbildungsmaßnahmen nutzbringend einzusetzen, sollen Arbeitslose nur dann solche Maßnahmen erhalten, wenn sie voraussichtlich dazu führen, wieder eine Arbeitsstelle zu erhalten. Kritiker der Software bemängeln, dass sie nicht den gesamten Quellcode einsehen durften und deshalb befürchten, die Software könne Frauen und ältere Menschen diskriminieren, indem sie ihnen geringere Chancen bescheinigt.

- Bei Strafverfahren in den USA wird eine Software namens Compas eingesetzt, um die Rückfallgefahr und Gewaltbereitschaft verurteilter Straftäter zu ermitteln. Davon abhängig legt der Richter Maßnahmen während und nach der Haftdauer fest, die das Rückfallrisiko vermindern sollen. Etliche Medienberichte, etwa (Ziegler 2017), deuten an, diese Software habe Einfluss auf die Haftdauer. Das ist falsch, wie in (Wisconsin 2016) ausführlich dargestellt. Dennoch verbleibt Kritik an der Software, weil sie etwa dunkelhäutigen Menschen schlechtere Prognosen ausstellt als anderen, und das, obwohl die Hautfarbe nicht zu den Eingabedaten gehört.

- Ein klassisches Fallbeispiel ist die maschinelle Prüfung der Bonität von Antragstellern für einen Kredit. Die vorherrschende Befürchtung lautet, diese Systeme würden etwa die Kommunikationsplattformen im Internet durchsuchen und Leuten mit auffälligem Verhalten den Kredit verweigern. (O'Neil 2016, 198..199) äußert eine originellere Kritik. Sie lehnt es ab, etwa die Wohnanschrift der Antragsteller als Indikator für die Bonität zu verwenden, und das, obwohl sie erfahrungsgemäß ein guter Indikator ist. Sie bezeichnet es als ungerecht, einen Menschen nach dem Verhalten anderer zu beurteilen, nur weil diese ähnliche Eigenschaften haben wie er selbst. Gerecht sei ausschließlich, einen Menschen nach seinem eigenen Verhalten zu beurteilen.

## Ethik, Moral und Software-Engineering

Kant hat die Philosophie in vier Teilgebiete eingeteilt und jedes mit einer Leitfrage versehen. Die Ethik ist eines davon, ihre Leitfrage lautet:

*Was soll ich tun?*

Sie enthält zwei entscheidende Wörtchen:

- *tun:* Ethik ist eine Handlungswissenschaft. Es gibt eine Vielzahl von Formaten, die zu unserem Thema erklären, was wir befürchten oder uns wünschen. Der Job des Ethikers ist aber erst dann erledigt, wenn klar ist, was getan werden muss, um einen erwünschten Zustand zu erreichen.

- *ich:* Gemeint ist hier natürlich nicht eine bestimmte Person, vielmehr eine Rolle, die Gesamtheit aller Personen mit einer bestimmten Aufgabe, Verantwortung oder in einer bestimmten Situation.

Hier ist schon sprachlich die Nähe zum Software-Engineering zu erkennen, denn auch dieses gibt Antworten zu Fragen dieser Form. Wie kommt ein Requirements-Engineer zu einer guten Vorgabe? Wie verständigen sich Tester, Entwickler und Management über die Folgen eines fehlgeschlagenen Tests?

Eine **Moral** ist eine Sammlung von Handlungsurteilen, zustimmenden oder ablehnenden. **Ethik** ist die Wissenschaft über Moral; sie verhält sich zur Moral also ähnlich wie die Physik zum Ingenieurwesen.

## Tätigkeiten und Rollen bei der Softwareentwicklung

Meine Überlegungen sind unabhängig von der Aufbau- und Ablauforganisation des Prozesses. Fast überall wird zwar behauptet, agil vorzugehen; man verwendet Begriffe wie Sprint, User Story und Product Owner. Die konkrete Ausformung variiert jedoch beträchtlich. Die folgende Aufstellung ist deshalb keine neue Entwicklungsmethode, sondern eine Abstraktion über alle solche Methoden *(siehe Abbildung 1)*. Die genannten Tätigkeiten müssen in jedem Fall ausgeführt werden, mal mehr oder weniger häufig oder intensiv. Wer immer gerade eine solche Tätigkeit ausführt, ist Träger der bei ihr genannten Rolle.

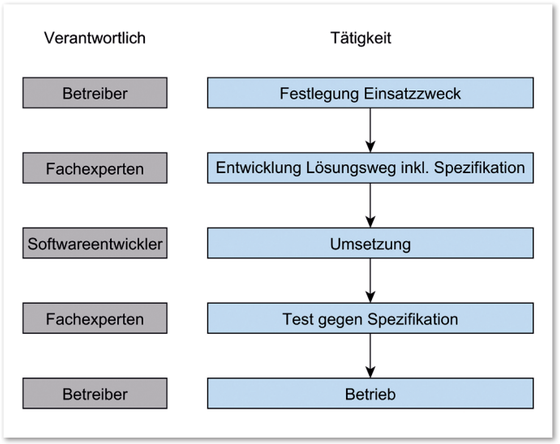


*Abbildung 1: Rollen und Tätigkeiten bei der Softwareentwicklung (© Thomas Matzner)*

Die erste entscheidende Frage beim Einsatz von Informatiksystemen ist ihr **Einsatzzweck**. Wer immer entscheidet, für einen bestimmten Zweck ein bestimmtes Verfahren einzusetzen, mit oder ohne Hilfe von Maschinen, ist **Betreiber** dieses Verfahrens. Oft ist dies die Leitung ei-

nes Unternehmens, einer Behörde oder einer sonstigen Organisation, aber nicht immer. Richter genießen eine hohe Unabhängigkeit. Wenn also etwa ein Richter entscheidet, zur Klärung eines bestimmten Sachverhalts einen Sachverständigen, eine Glaskugel oder eben eine Software einzusetzen, ist er in der Betreiberrolle.

Der Einsatzzweck soll möglichst konkret und nachprüfbar angegeben werden, also nicht nur „verbesserte Zuteilung von Fördermaßnahmen", sondern etwa in die Richtung: „Erhöhung der Vermittlungsquote geförderter Arbeitsloser von derzeit 60 auf 80 Prozent." Doch auch ein präziser Einsatzzweck lässt sich oft nicht direkt in Software umsetzen. Ein Softwareentwickler, der diese Zielvorgabe bekommt, würde zu Recht einwenden, keinen Sachverstand zu besitzen, um die Vermittlungschancen von Arbeitslosen zu beurteilen. Er würde eine **Spezifikation** fordern, die von **Fachexperten** auf der Fachdomäne erarbeitet wird, der der Einsatzzweck angehört. Genau betrachtet, braucht man häufig nicht nur diese Spezifikation, die das Verhalten der Software bestimmt. Das **Verfahren**, um den Einsatzzweck zu erreichen, kann neben maschinellen auch menschliche Tätigkeiten enthalten; die Beschreibung des Verfahrens, die auch die Spezifikation enthält, nenne ich den **Lösungsweg**.

Über die Umsetzung durch Softwareentwickler muss ich nicht viele Worte verlieren. Wichtig ist, dass der T**est gegen die Spezifikation**, hinreichend gründlich durchgeführt, Mängel in der Umsetzung erkennen lässt. Da der Betreiber, wie oben gefordert, die Verantwortung für den Einsatz des Verfahrens hat, hat er das Recht, aber auch die Verpflichtung, auf Basis des Testergebnisses zu entscheiden, ob er den Einsatz verantworten kann.

## Einsatzzweck vs. Verfahren

Ein grundlegendes Prinzip lässt sich jetzt schon erkennen. Die moralische Bewertung des Einsatzzwecks lässt sich von der für das eingesetzte Verfahren trennen. Fragen nach den Chancen von Arbeitslosen, der Gefährlichkeit von Straftätern und der Kreditwürdigkeit wurden auch schon gestellt, bevor es Computer gab. Viele Formate zu unserem Thema vermischen das. Sie äußern Befürchtungen über – absichtlich oder nicht – schlecht gemachte Software, um dann Kritik etwa am kapitalistischen Wirtschaften, der Strafjustiz oder dem Sozialstaat zu üben. Eine konstruktive Lösung kommt umgekehrt zustande: Solange wir keine Einigkeit über die moralische Zulässigkeit des Einsatzzwecks haben, brauchen wir über unterschiedliche Lösungswege gar nicht zu sprechen.

> *Die Verantwortung für den Einsatzzweck trägt der Betreiber, womöglich langfristig unterstützt durch staatliche und gesellschaftliche Organe, die gesetzliche oder sonstige Moralvorschriften setzen.*

> *Die Verantwortung der anderen Beteiligten beschränkt sich darauf, den Einsatzzweck durch ein Verfahren möglichst unverfälscht zu unterstützen.*

Konkret: Warum rationiert man überhaupt die Fördermittel für Arbeitslose und gibt nicht allen einfach das, was sie sich wünschen? Das ist das Grunddilemma jedes Sozialstaates, denn das, was er an die aktuell Bedürftigen ausschüttet, muss er den aktuell Erfolgreichen abnehmen. Derselbe Arbeitslose, dem heute eine Fortbildung verweigert wird, hat fünf Jahre zuvor über die hohe Beitragslast gestöhnt. Für dieses Problem gibt es keine Patentlösung; es zu lösen, ist jedenfalls nicht unsere Kompetenz als Softwareentwickler. Selbst, wenn das ganze Informatikpersonal rund um das Arbeitsministerium sich zusammentäte und eine ihm gerecht erscheinende Umverteilungsmethode implementieren würde, wäre das undemokratisch. Nicht die Softwareentwickler haben den Auftrag, politische Fragen zu entscheiden, sondern über freie Wahlen letztlich der Arbeitsminister. Wer sich bei uns über ungerechte Verteilung der Mittel beschwert, dem ist zu sagen, er soll nächstes Mal andere Politiker wählen.

Überall hört und liest man über die Allgegenwart von Informatiksystemen. Tatsächlich sind wir in vielen, wenn auch nicht allen, Lebensbereichen mit unseren Systemen *unterstützend* tätig. Gerade deshalb ist es wichtig, nicht den Anspruch zu erheben, in allen Fachdomänen *mitzubestimmen*. Ein Student erzählte mir von einem Professor, der in einem Vortrag den Spruch brachte: „Die Informatik frisst die Welt." Genau das darf nicht passieren, denn „die Informatik" ist im Sinn der Ethik kein Akteur, dem man Verantwortung zuschreiben, Regeln auflegen und Verstöße sanktionieren kann. Es ist wichtig, die Betreiber der Systeme immer wieder darauf hinzuweisen, dass sie zwar nervtötende Rechenschritte an die Maschine abgeben können, nicht jedoch ihre Verantwortung und das Verständnis für ihre Domäne, das sie brauchen, um sie wahrzunehmen.

## Voraussage vs. Entscheidung

Die eingangs vorgestellten Fälle decken natürlich nicht das gesamte Feld von Informatiksystemen ab, aber einen Ausschnitt, der offensichtlich besondere Besorgnis erregt. Sie haben ihre Grundstruktur gemeinsam: Sie bestehen aus einem Voraussage- und einem darauffolgenden Entscheidungsschritt. Auf die Voraussage (Wird ein geförderter Arbeitsloser wieder einen Job finden? Ein Straftäter rückfällig werden? Ein Kreditnehmer den Kredit vollständig bedienen?) folgt eine Entscheidung. Wir kennen das Entwurfsprinzip „divide and conquer", um schwer überschaubare Probleme auf handhabbare Größe herunterzubrechen. Trennung von Einsatzzweck und Verfahren war ein solcher Schritt, Trennung von Voraussage und Entscheidung ist der nächste. Welche Erleichterung bringt er uns?

> *In vielen Fällen ist der Voraussageschritt moralisch unkritisch, jedoch fachlich schwierig.*

> *In vielen Fällen ist der Entscheidungsschritt moralisch kritisch, jedoch fachlich einfach.*

Sehen wir uns die Kreditentscheidung an: Eine möglichst zutreffende Voraussage nützt bei vernunftorientierter Betrachtung beiden Parteien. Dem Kreditgeber sowieso, der gute von schlechten Risiken trennen kann. Aber auch für den Kreditnehmer ist eine negative Voraussage, vorausgesetzt sie ist zuverlässig, zwar momentan unangenehm, aber in Summe nützlich, bewahrt sie ihn doch womöglich von den unangenehmen Folgen der Überschuldung. Dass die Voraussage jedoch fachlich und in der Folge auch in der technischen Umsetzung schwierig ist, dürfte auf der Hand liegen. Die rein finanziellen Voraussetzungen, etwa Vermögen, Einkünfte, wirtschaftliche Lage des Berufsstandes, des Unternehmens und der Branche des Kreditnehmers, mögen noch fassbar sein. Aber wie sieht es mit menschlichen Risiken aus, mit Krankheiten, Auseinanderbrechen von Familien und sonstigen Schicksalsereignissen? Der Umgang mit solchen Unsicherheiten wird uns noch beschäftigen müssen.

Nun zum Entscheidungsschritt. Würde der Voraussageschritt in jedem Einzelfall ein klares Ja-Nein-Ergebnis liefern, wäre er trivial. Das können wir aber nicht erwarten, allenfalls eine Wahrscheinlichkeit für das jeweils erwünschte Ereignis. Der Banker könnte es sich nun einfach machen und eine betriebswirtschaftliche Rechnung darüber anstellen, wie hoch die Wahrscheinlichkeit auf Rückzahlung sein muss, um in Summe noch ein gutes Geschäft zu machen. Soll er bei 80 Prozent den Kredit vergeben? In vier von fünf Fällen macht er satten Gewinn, auch in dem fünften Fall verliert er nicht alles, sondern nur einen Teil der Summe. Das mag sich für ihn noch lohnen, lässt allerdings die Risiken für den Kreditnehmer außer Acht. Dieser ist im Fall der Zahlungsunfähigkeit wahrscheinlich überschuldet und muss das Verfahren der Privatinsolvenz über sich ergehen lassen. Dies bringt mit sich, über Jahre hinweg von allen erwirtschafteten Einkünften nur das gesetzlich festgelegte Existenzminimum behalten zu dürfen und alles darüber hinaus an die Gläubiger zu geben. Dieses Existenzminimum ist so festgelegt, dass es gerade ausreicht, ein menschenwürdiges Leben zu führen. Ein leichtfertig vergebener Kredit bringt den Kreditnehmer also zu weitgehendem Verlust oder der Einschränkung seiner Menschenrechte. Diese Folge überschattet alle anderen, die keinen Menschenrechtscharakter haben; also ist geboten, ihn davor bestmöglich zu schützen und dafür auch das Gewinnstreben der Bank zu begrenzen.

Wie hoch soll nun die Wahrscheinlichkeit für korrekte Rückzahlung sein, um den Kredit verantworten zu können? 100 Prozent werden wir angesichts der aufgezählten unkalkulierbaren Risiken nicht erreichen. Genügen 80 oder sollten wir 90 fordern? Auf **Grenzziehungen** wie diese gibt auch die Ethikliteratur keine praktisch verwertbare Antwort.

Hier hatten wir es wenigstens nur mit *einem* Menschenrecht zu tun, also einem klaren Fokus der Interessen. Beim Straftäter ist es noch komplizierter. Nehmen wir einen gefährlichen Fall an, etwa Misshandlung von Kindern. Bei solchen Tätern tun sich nach Abbüßung der Strafe und Therapie auch Fachleute schwer, eine Prognose für die Rückfallgefahr abzugeben, die als Voraussetzung dafür dient, dem Täter wieder die volle Freiheit zu gewähren. 80 Prozent erscheinen hier zu niedrig, bedeutet das doch, in einem von fünf Fällen wieder ein Kind zu schädigen. Aber wieder müssen wir irgendeine Grenze ansetzen. Nehmen wir an, sie liege bei 99 Prozent. Dann bedeutet das, dass in der Täterkohorte mit nur 98 Prozent Erfolgswahrscheinlichkeit 98 von 100 Tätern ihrer Freiheit beraubt werden, obwohl sie inzwischen ungefährlich sind. Auch dies ist ein schwerwiegender Eingriff in die Menschenrechte und läuft der Idee des Rechtsstaates, dass eine Tat schlussendlich verbüßt sein kann, zuwider.

Diese Beispiele sollten auch illustriert haben, dass Erwägungen wie diese nicht Sache des Informatikpersonals sein können. Sie zeigen auch einen Unterschied zwischen Ethik und Software-Engineering, nachdem ich anfangs eine Gemeinsamkeit gezeigt hatte. Wenn man etwa Modellierung lehrt, kann man am Ende jeder Unterrichtseinheit die Teilnehmer mit einer mustergültigen Lösung nach Hause schicken, über die sich alle freuen können. Für viele alltägliche Aufgaben der Ethik gibt es keine solche Lösung, über die wir in Jubel ausbrechen können. In manchen Fällen lässt sich der Konflikt abmildern, indem man statt einer Ja-Nein-Entscheidung differenziertere Maßnahmen einsetzt. So gibt es für Straftäter mit unklarer Prognose nicht nur die Möglichkeit der lebenslangen Sicherungsverwahrung, sondern abgestufte Maßnahmen der Überwachung nach der Haftentlassung. Die Kreditentscheidung ist hier unerbittlich: Ist der Kredit einmal vergeben, gibt es für alle Beteiligten nur mehr geringe Spielräume, um von den dadurch gesetzten Regeln abzuweichen.

Ist diese moralisch schwierige Entscheidung, bei welchem Voraussageergebnis welche Folge eintreten soll, einmal getroffen, ist ihre Umsetzung denkbar einfach. Ob man den Vergleich mit einem Schwellwert durch einen Menschen oder eine Maschine ausführen lässt, ist unerheblich.

## Gerechtigkeit

Dies ist ein sperriges Konzept, von so vielen teils logisch widersprüchlichen Forderungen durchdrungen, dass es kaum möglich ist, für eine Aufgabe eine nachweisbar gerechte Lösung zu finden. Sehen wir uns wieder das Kreditwesen an. Basis dafür ist unser Schuldrecht, das unter dem Stichwort der Privatinsolvenz die Risiken zwischen Kreditgeber und Kreditnehmer verteilt: Der Kreditnehmer erleidet während deren Dauer erhebliche Nachteile, der Kreditgeber verzichtet danach auf die noch bestehende Restforderung. Dies führt dazu, dass beide bestrebt sind, Ausfälle zu vermeiden.

Das ist ungerecht, denn es führt dazu, dass Reiche immer reicher, Arme zumindest schwieriger reich werden. Wer schon zwei Mietshäuser besitzt, kann der Bank diese als Sicherheit anbieten und hat satte Mieteinnahmen, erhält also leicht einen Kredit für ein drittes Mietshaus. Alle drei vererbt er seinen Kindern, die ein zumindest finanziell sorgloses Leben führen können. Wer kein Kapital und nur ein durchschnittliches Arbeitseinkommen hat, bekommt womöglich keinen Kredit, vererbt seinen Kindern nichts, und alle müssen für ihren Lebensunterhalt arbeiten. Dies zu ändern, würde gravierende Eingriffe in das Schuldrecht erfordern, für die kaum ein praktikabler Ansatz erkennbar ist. Jede Entlastung einer der beiden Parteien würde zu einer Belastung der anderen führen und deren Bereitschaft mindern, das Geschäft einzugehen. Ein Teufelskreis.

Doch der anfangs zitierte Einwand gegen die Ermittlung von Indikatoren für die Kreditwürdigkeit zielte nicht auf diese grundsätzliche Ungerechtigkeit, sondern auf das Verfahren im konkreten Einzelfall. Was kennzeichnet eine gerechte Bonitätsprüfung?

Bei aller Sperrigkeit gibt es eine einfache Regel für gerechtes Handeln, nämlich Gleiches gleich zu behandeln. Wird einem rückzahlungsfähigen Kunden der Kredit verwehrt, ist das ungerecht, denn ihm wird im Gegensatz zu den anderen eine Chance verwehrt. Wird er einem nicht rückzahlungsfähigen gegeben, ist das auch ungerecht, denn er wird im Gegensatz zu den anderen nicht vor der Überschuldung geschützt.

Wir sehen also, dass zumindest in den Fällen, in denen kein Interessenkonflikt besteht, das gerechteste Verfahren dasjenige ist, das möglichst wenig Fehler macht. Das zweckmäßigste Verfahren ist gleichzeitig das gerechteste. Verfahren, die keine Trefferrate von 100 Prozent erzielen können, sind deshalb niemals vollständig gerecht, aber je besser die Quote, desto gerechter ist auch das Verfahren. Man beachte, dass für diesen Schluss kein Einblick in die inneren Abläufe des Verfahrens nötig war, sondern lediglich die Betrachtung seiner Außenwirkung, was uns zum nächsten Thema bringt.

## Innen- vs. Außenwirkung, Bedeutung des Testens

Wenn ich Texte lese, in denen vor den Gefahren eines Informatiksystems gewarnt wird, weil man nicht weiß, welche schädlichen Wirkungen es entfaltet, frage ich mich immer: Warum hat niemand dem Autor erklärt, dass man diese Systeme testen kann und muss? Oder, wenn er es weiß, warum erklärt er es nicht seinen Lesern als effektives Mittel, die Befürchtungen zu entkräften?

Eine Einschränkung vorweg: Natürlich kann der Bankkunde, der Arbeitslose oder Straffällige ein Misstrauen gegenüber der Großorganisation empfinden, der er gegenübersteht. Er hat nicht die Möglichkeit, die dort eingesetzte Software systematischen Tests zu unterziehen; bestenfalls kennt er einen Testfall, seinen eigenen, den er nicht mit anderen vergleichen und etwa auf Gerechtigkeit prüfen kann. Dies ist jedoch keine Spezialität von Informatiksystemen, sondern der Moralregeln für die Wahrung der Grenzen zwischen Personen, seien es natürliche oder juristische. Auch bei rein menschlicher Abwicklung hätten die Genannten keinen Einblick in die Abläufe bei Bank, Arbeitsamt oder Gericht gehabt. Am transparentesten ist noch das Gericht, das öffentlich verhandelt und schriftlich begründet. Aber wie ist der Richter zu der Überzeugung gekommen, der Angeklagte sei weiterhin gefährlich und nach Entlassung mit einer elektronischen Fußfessel zu überwachen? Vielleicht hat er selbst gründlich nachgedacht, vielleicht seine Frau oder seinen alten Hochschullehrer angerufen, vielleicht auch eine Software rechnen lassen – wir werden es nicht erfahren. Wenn wir also von Tests sprechen, setzt das voraus, dass der Betreiber selbst ein Interesse an einem guten Ergebnis hat oder dass, wie bei Jahresabschlüssen oder bestimmten technischen Anlagen schon üblich, eine Prüfinstanz stellvertretend für den Rest der Bevölkerung die Abläufe bewertet.

Bei der Aufstellung einer Informatikethik benutze ich vorwiegend die Werkzeuge der **konsequentialistischen Ethik**, die Handlungs*folgen* bewertet, etwa im Gegensatz zu Tugend- oder Gesinnungsethik. Die Konzentration auf Folgen macht deutlich, dass es nicht auf die inneren Abläufe der Handlung ankommt, sofern sie die erwünschten Folgen erzeugt. Also kommt es innerhalb des Verfahrens auch nicht auf den inneren Aufbau und Abläufe der Software an, sondern auf die Folgen von deren Ausführung, also auf die Außensicht. Genau diese können wir durch Testen ermitteln. Damit erreichen wir zwar nicht die Sicherheit eines mathematischen Beweises, können jedoch den Test so intensiv gestalten, wie es der Kritikalität der Anwendung angemessen ist.

> *Die Kritikalität eines Einsatzzwecks misst sich an dem maximal möglichen immateriellen und materiellen Schaden, der bei seiner Verfehlung eintreten kann.*
>
> *Die Intensität des Tests einer Software muss der Kritikalität des Einsatzzwecks angemessen sein. Verantwortlich hierfür ist der Betreiber, der das Testergebnis für die Einsatzentscheidung braucht.*

Die Kritiker der Arbeitsamts-Software wollten sich erst mit Einblick in den Quellcode beruhigen lassen. Es gibt mancherlei Gründe, Quellcode offenzulegen, allerdings für manche Softwarehersteller auch Gründe dagegen. Für die Überprüfung der moralischen Integrität der Software ist der Test zu bevorzugen:

- Die Überprüfung des Quellcodes müsste durch eine von den ursprünglichen Entwicklern unabhängige Instanz geschehen, da die Entwickler Urheber des moralischen Problems sein könnten. Da in diesem Fall das Fehlverhalten gut versteckt sein könnte, muss die Prüfinstanz eine ihr unbekannte Software in ihrer Gänze überprüfen, was hohe Aufwände verursacht.
- Beim Lesen von Programmcode unterliegen wir kognitiven Verzerrungen, die uns hinsichtlich der Funktionsweise in die Irre führen. Gerade deshalb müssen wir ja testen.
- Die Ergebnisse eines Tests sprechen die Sprache des Betreibers, der die Einsatzentscheidung treffen muss.

Illustrieren wir das an der Frage, ob die Arbeitsamts-Software Frauen benachteiligt. Will man das am Code überprüfen, muss man den gesamten Code inspizieren. Man darf sich nicht darauf verlassen, dass eine eventuelle Ungleichbehandlung von Frauen in dem zentralen Modul zur Überprüfung der Chancen lokalisiert ist, auch nicht darauf, dass die entsprechenden Datenfelder an der Stelle semantisch korrekt mit „Geschlecht" oder Ähnlichem benannt sind. Also ist zeitraubende Detektivarbeit am gesamten Code angesagt. Ist nichts gefunden, könnte man es immer noch übersehen haben. Ist etwas gefunden, kann immer noch unklar sein, wie häufig sich der Unterschied auswirkt und ob er tatsächlich unerwünschte Ergebnisse liefert.

Ein Test hingegen kann folgendermaßen ablaufen: Man lässt zuerst die Eingabedaten für die Arbeitslosen der letzten zwei Jahre durchlaufen und hält die Ergebnisse fest. Nun ändert man wahllos in den Eingabedaten das Geschlecht und zur Sicherheit den Vornamen, falls da eine Plausibilitätsprüfung zuschlägt. Man lässt die Daten wieder durchlaufen. Ist das Ergebnis unverändert, hat man zwar keine absolute Sicherheit, kann aber sagen, dass immerhin in den letzten zwei Jahren keine Diskriminierung vorgekommen wäre. Gibt es Abweichungen, kann der Betreiber sich die Fälle zeigen lassen und entscheiden, ob sie sinnvoll sind. Wenn etwa Frauen seltener zu Straßenbauarbeiterinnen umgeschult werden sollen als Männer, kann der Arbeitsminister damit vermutlich leben. Wenn aber Frauen systematisch zu schlechter bezahlten Jobs umgeschult werden, ist das ein Alarmsignal.

## Ist der Lösungsweg korrekt?

Ein Test gegen die Spezifikation kann Fehler in ihr nicht aufdecken. Wie sicher können wir uns des Lösungswegs und der in ihm enthaltenen Spezifikation sein? Das hängt von der Fachdomäne ab, der der Einsatzzweck angehört *(siehe Abbildung 2)*.

Es gibt Fachdomänen mit nachweisbar gültigen Regeln, also solchen, die nicht nur von einigen Personen für richtig gehalten werden, sondern die intersubjektiv beweisbar sind. Dazu gehören die exakten Wissenschaften und darauf aufbauend technische Produkte und Vorgänge. Es gibt auch Domänen, in denen die Menschen die Macht haben, solche Regeln als gültig festzulegen, etwa wirtschaftliche und administrative Transaktionen. Jemand bietet eine Ware an, jemand bestellt sie, sie wird geliefert, eine Rechnung geschrieben, Geld- und Warenfluss verbucht, Steuern darauf erhoben – das alles läuft nach vorbestimmten, zweifelsfrei gültigen Regeln.

In anderen Fachdomänen gibt es solche Regeln nicht. Dazu gehört menschliches Verhalten. Kein Wirtschafts-Nobelpreisträger, der ein mathematisches Wirtschaftsmodell entwickelt hat, kann mit Sicherheit den morgigen Kurs einer Aktie voraussagen, geschweige

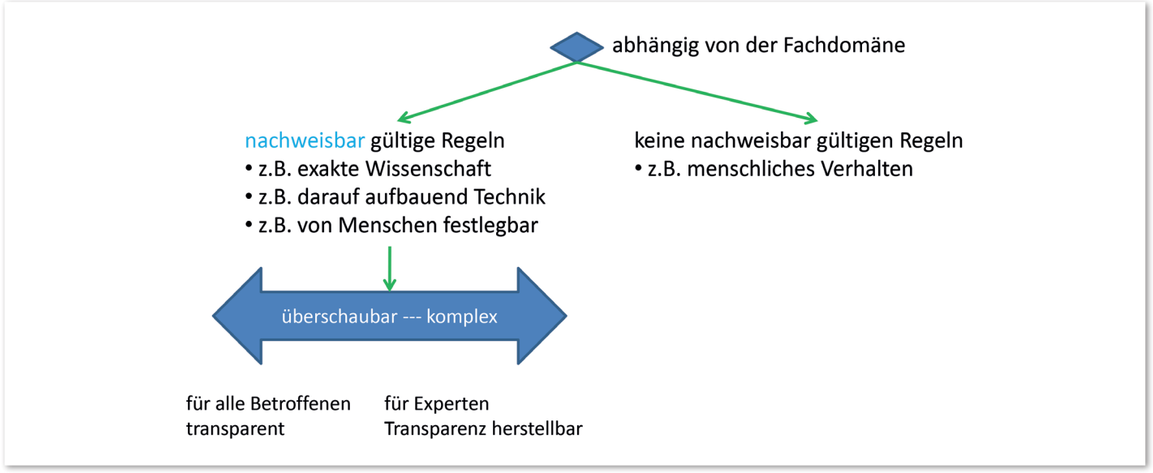


*Abbildung 2: Klassifikation des Einsatzzwecks (© Thomas Matzner)*

denn langfristige menschliche Entwicklungen wie in unseren Fallbeispielen. Auch die psychologische Literatur entwickelt zwar Handlungsmuster, aber nur für kurzfristig ablaufende Vorgänge wie etwa die Nahrungs- oder Partnersuche. Wir müssen uns damit abfinden, dass es für viele, auch alltägliche, Aufgaben keinen offensichtlich gültigen Lösungsweg gibt *(siehe Abbildung 3)*.

Hat man nachweisbar gültige Regeln, setzt man diese im Lösungsweg einfach um. Je nach Komplexität kann auch dies schwierig sein, so ist etwa die Rechnungsschreibung einfacher als die Ermittlung von Steuern.

Gibt es solche Regeln nicht, muss man sich anders behelfen. Man kann annähernde Regeln verwenden, bei der Kreditprüfung etwa Grenzen für Einkünfte und Startkapital festlegen. Will man aber etwa die gesundheitliche und familiäre Entwicklung des Kreditnehmers beurteilen, wird es schon schwieriger. Datengetriebene Verfahren, zu denen die meisten Techniken der künstlichen Intelligenz zählen, gehen einen anderen Weg: Sie ermitteln Korrelationen zwischen Ausgangsdaten und erwünschtem Ergebnis, etwa der tatsächlichen Rückzahlung des Kredits. Während man diese beiden Lösungswege ganz oder teilweise maschinell abwickeln kann, ist bei rein menschlicher Abwicklung noch ein dritter denkbar: das Vertrauen auf menschliche Intuition, bevorzugt durch Experten mit langjähriger einschlägiger Erfahrung.

Einsatzzwecke ohne nachweisbar gültige Regeln oder mit so komplexen Regeln, dass deren Umsetzung in einen Lösungsweg schwierig und fehlerträchtig ist, nenne ich **schwierig algorithmisierbar**.

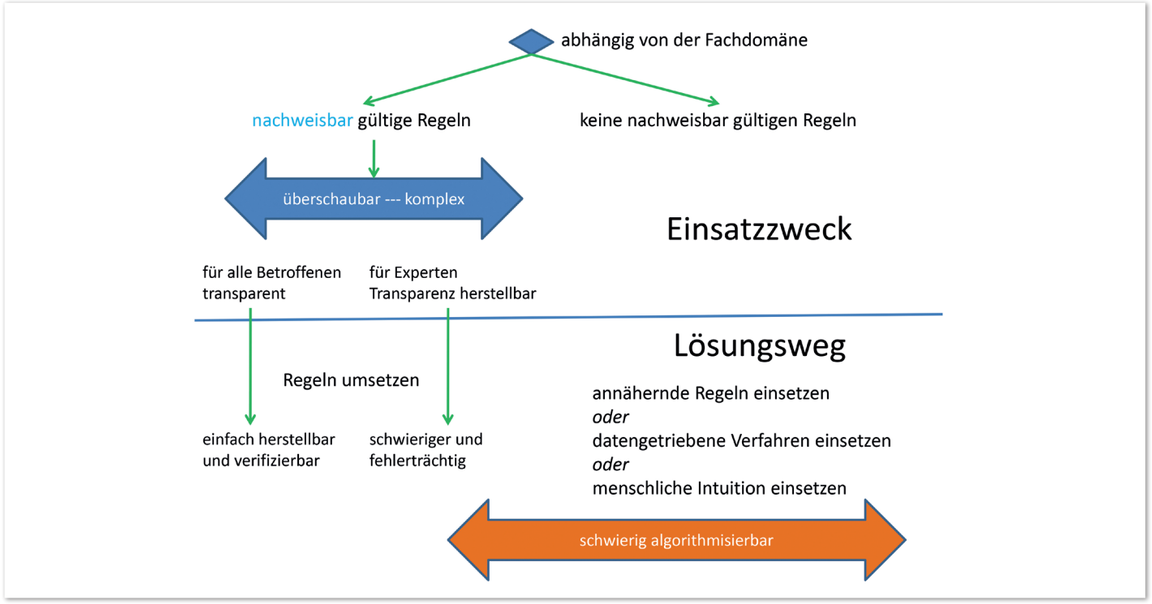


*Abbildung 3: Mögliche Lösungswege abhängig vom Einsatzzweck (© Thomas Matzner)*

Strenggenommen geht es hier nicht um die *Formulierung* eines Algorithmus, die man sich beliebig einfach machen kann, sondern um den Nachweis, dass Lösungsweg und Spezifikation tatsächlich den Einsatzzweck treffen.

Wie schon bei der Diskussion von Voraussagen erwähnt, ist in diesen Fällen ein *korrekter* Lösungsweg nicht zu erwarten, also einer, der in jedem Einzelfall das richtige Ergebnis liefert. Das Beste, was wir verlangen können, ist ein möglichst *verlässlicher* Lösungsweg.

> *Bei einem schwierig algorithmisierbaren Einsatzzweck ist von allen Lösungswegen der verlässlichste zu wählen, der die niedrigste Fehlerrate liefert.*

Das bringt uns zu der Frage, wie man diesen findet, wenn man seine Verlässlichkeit nicht direkt aus seiner Formulierung ableiten kann.

## Test gegen die Wirklichkeit

Es gibt schon seit Längerem Informatiksysteme, die nicht nur gegen ihre Spezifikation getestet werden. Denken wir an vollautomatisch fahrende Kraftfahrzeuge. Es gibt sie schon seit Jahren auf unseren Straßen, aber nicht im Einsatz für normale Konsumenten. In ihnen sitzt ein Begleitfahrer, der das Verhalten des Fahrzeugs überwacht und im Notfall eingreifen kann. Wenn dieser Vorgang genügend lange durchgeführt wurde, um Vertrauen zu dem System haben zu können, kann es als praxisreif gelten. Dieses Vorgehen nenne ich **Test gegen die Wirklichkeit**, denn hier wird nicht gegen eine vorbestimmte Spezifikation auf Papier getestet, sondern gegen eine nicht immer vorbestimmbare Wirklichkeit *(siehe Abbildung 4)*.

Während also der Test gegen die Spezifikation Fehler bei der Umsetzung aufdeckt, leistet der Test gegen die Wirklichkeit dies für Lösungsweg und Spezifikation.

Der vorhin beschriebene Test auf Gender-Ungerechtigkeit war kein Test gegen eine Spezifikation, sondern gegen die Wirklichkeit: Wir hatten bei ihm kein Ergebnis fest vorgegeben, sondern erforscht, welche Wirkung eine angenommene Vertauschung der Geschlechter hätte.

Damit können wir die Forderung aus dem letzten Abschnitt konkretisieren:

> *Bei einem schwierig algorithmisierbaren Einsatzzweck ist durch Test gegen die Wirklichkeit zu ermitteln, welcher Lösungsweg die niedrigste Fehlerrate liefert.*

Diese Forderung ist zugegebenermaßen zu streng: Sie würde bedeuten, *alle* möglichen Lösungswege zu implementieren und gründlich gegen die Wirklichkeit zu testen. Das kann etwa eine einzelne Bank, wie reich sie auch sein möge, kaum leisten. Es müssten alle annähernden Regelsysteme, alle bisher entwickelten datengetriebenen Algorithmen und eine hinreichend große Anzahl von Menschen gegeneinander getestet werden. Was man aber in der Praxis verlangen kann, ist mindestens, für das derzeit bevorzugte Verfahren einen solchen Test durchzuführen sowie Verfahren, für die das mit wenig Aufwand möglich ist, etwa auf dem Markt angebotene fertige Software, damit zu vergleichen.

## Mensch vs. Maschine

In manchen ethischen Leitlinien wird gefordert, dass bestimmte kritische Aufgaben, wie sie in unseren Fallstudien vorkommen, grundsätzlich oder auf Verlangen des Betroffenen von Menschen gelöst werden. Daraus könnte man schließen, dass Menschen diese Aufgaben zuverlässig besser lösen als Maschinen. Tatsächlich haben wir uns über Jahrtausende daran gewöhnt, dass Lehrer ihre Schüler beurteilen, Richter den Charakter der Prozessbeteiligten und Vorgesetzte die Leistung ihrer Mitarbeiter. Ist also die maschinelle Ausführung oder Unterstützung bei dieser Aufgabe nur eine Notlösung, um Kosten oder Zeit zu sparen?

(Kahneman 2011, 222..233) lässt uns daran zweifeln, dass menschliche Voraussagen von Ereignissen unter Unsicherheit der Goldstandard sind. Schon Mitte des vorigen Jahrhunderts, als noch kaum jemand an Computereinsatz dachte, war durch kontrollierte Experimente erwiesen, dass selbst eine einfache Scoring-Formel mindestens ebenso verlässliche Ergebnisse liefert wie intuitive Entscheidungen durch Experten. Diese Erkenntnisse laufen unserem Selbstbild, vor allem wenn wir berufs- und lebenserfahrene Experten sind, zuwider und werden selten in der Praxis thematisiert,

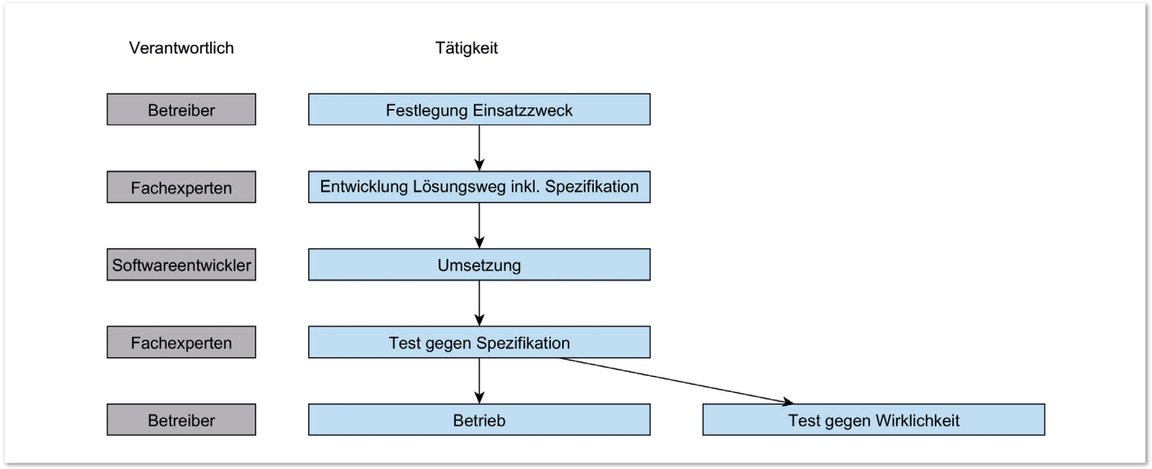


*Abbildung 4: Softwareentwicklung mit Test gegen die Wirklichkeit (© Thomas Matzner)*



obwohl sich das zitierte Buch gut verkauft hat. Als Ethiker habe ich hierzu nichts Eigenes beizutragen, muss aber die relevante Forschung zur Kenntnis nehmen und daraus schließen:

> *Bei einem Einsatzzweck, der Voraussagen unter Unsicherheit umfasst, darf nicht vorausgesetzt werden, dass ein rein durch Menschen ausgeführter Lösungsweg der verlässlichste ist.*

Das heißt nicht, dass der Mensch den Wettbewerb mit anderen Lösungswegen stets verliert, wohl aber, dass er sich ihm stellen muss, zumal es „den Menschen" gar nicht gibt, sondern viele unterschiedliche, deren Einschätzungen oft genug widersprüchlich sind. Die wichtigsten Schwächen der drei Lösungsweg-Typen für schwierig algorithmisierbare Aufgaben sind:

- Annähernde Regeln
  - Da sie per Voraussetzung nicht nachweisbar sind, entspringen sie menschlicher Intuition, siehe dort.
- Datengetriebene Verfahren
  - Korrelation bedeutet nicht Kausalität. Also liefert ein solches Verfahren auch bei häufiger Ausführung keine zusätzlichen *Einsichten* in die Mechanismen der Fachdomäne. Demnach ist es anfällig gegenüber Veränderungen in der Wirklichkeit, mangelhaften Trainingsdaten und anderen Versäumnissen.
- Menschliche Intuition
  - Die zugrunde liegende Datenbasis ist naturgemäß kleiner als bei den anderen Lösungswegen, bei denen potenziell alle überhaupt verfügbaren Daten genutzt werden können.
  - Sowohl unsere Erinnerungen wie auch die Schlüsse, die wir aus ihnen ziehen, unterliegen vielfältigen kognitiven Verzerrungen. Davon handelt das soeben zitierte Buch von Kahneman.
  - Also sind die vielfach befürchteten Vorurteile vorwiegend bei intuitiven Voraussagen oder aus Intuition gewonnenen annähernden Regeln zu erwarten.

## Testbarkeit gegen die Wirklichkeit

Der Test gegen die Wirklichkeit ist je nach Einsatzzweck unterschiedlich schwierig. Vergleichen wir dazu das vollautomatisch fahrende Auto mit der Kreditentscheidung. Das Autofahren ist immens schwierig algorithmisierbar, umfasst es doch die korrekte Interpretation des Verkehrsgeschehens in Echtzeit inklusive Voraussagen über die Absichten der Verkehrsteilnehmer. Es ist aber relativ leicht gegen die Wirklichkeit testbar: Der Begleitfahrer kann, Kenntnis der Verkehrsregeln vorausgesetzt, in jeder Situation sofort sagen, ob sich das Auto korrekt verhalten hat.

Umgekehrt bei der Kreditentscheidung. Eine einfache Heuristik auf Basis der finanziellen Situation lässt sich in wenigen Tagen spezifizieren; damit hat man womöglich schon eine Lösung, die auch heute noch mancherorts produktiv eingesetzt wird. Aber wie gestaltet sich der Test gegen die Wirklichkeit? Wenn heute eine Entscheidung getroffen wird, ist erst in Jahren oder Jahrzehnten objektiv ermittelbar, ob sie richtig war.

Zu den Merkmalen Kritikalität und Algorithmisierbarkeit des Einsatzzwecks tritt also als drittes die **Testbarkeit gegen die Wirklichkeit**. Sie ist bei Voraussagen umso schwieriger, je länger der Voraussagezeitraum ist.

Aus diesem Grund ist in der vorangegangenen *Abbildung 4* der Test gegen die Wirklichkeit nicht einfach vor dem Betrieb, sondern parallel zu ihm eingezeichnet. Anders als der Test gegen die Spezifikation kann er nicht einfach vor Inbetriebnahme durchgeführt und dann abgehakt werden, sondern erfordert in schwierigen Fällen ständiges Nachprüfen. Hier einige zusätzliche Maßnahmen für solche schwierigen Fälle:

- In erster Linie denkt man natürlich bei langfristigen Voraussagen an die Verwendung historischer Daten. Man kann bei der Kreditentscheidung nicht abwarten, bis die Antwort für heute gewährte Kredite da ist, also geht man Jahre bis Jahrzehnte zurück. Man muss sich jedoch je nach Fachdomäne überlegen, wie gut historische Verhältnisse auf gegenwärtige und künftige übertragbar sind. So kann etwa eine Branche, die in den letzten Jahrzehnten glänzend florierte, zusammen mit den Regionen, in denen sie Arbeitsplätze bietet, zu guten Kreditverläufen führen. Wenn es der Branche künftig schlechter ergeht – man denke etwa an mögliche Verwerfungen im Zusammenhang mit der Energie- und Verkehrswende –, können historische Gewissheiten sich verflüchtigen.

- Deshalb sollte man während des Betriebs laufend überwachen, ob sich die Datenkonstellationen verändern, ob etwa bestimmte Muster mehr oder weniger häufig auftreten als bei den anfangs zugrunde gelegten Tests.

- Wenn mehrere Verfahren zur Verfügung stehen, etwa auch nur durch unterschiedliche Konfiguration desselben Grundverfahrens, kann es lohnen, diese parallel einzusetzen, um festzustellen, ob ein anderes als das aktuell entscheidende plötzlich Hinweise auf Veränderungen zeigt.

- Um all dies zu ermöglichen, müssen in den unterstützten Prozessen laufend Erfolgsdaten zuvor getroffener Entscheidungen eingespielt werden. Das ist derzeit nicht in allen Fällen selbstverständlich. Natürlich hat die Bank aus ihrem operativen Geschäft jederzeit vollständige Daten über die Bedienung von Krediten. Aber hat jeder Richter solche Daten über die Rückfallquote von Straftätern mit bestimmten Merkmalen zur Verfügung? Bei der Erfassung und Verarbeitung solcher Daten könnten Einwände hinsichtlich des Datenschutzes gemacht werden. Die Darstellung des moralischen Dilemmas beim Umgang mit therapierten Missbrauchstätern hat jedoch den Blick auf Menschenrechte geweitet. Es gibt nicht das eine Recht auf Schutz privater Daten, dem alle anderen Interessen unterzuordnen sind. Zu den Grundaufgaben der Angewandten Ethik gehört das ständige Austarieren solcher Werte ohne vorgefertigte Präferenzen.

Ein moralisches Dilemma bei schwierig algorithmisier- und testbaren Einsatzzwecken entsteht durch selbsterfüllende Prophezeiungen. Ist ein Kredit verweigert oder ein Straftäter unter dauerhafte Überwachung gestellt worden, kann sich der Akteur immer in der wohligen Überzeugung wiegen, großen Schaden verhindert zu haben. Ein objektives Feedback aus der Wirklichkeit kann nicht stattfinden, da ja sowohl das befürchtete Ereignis als auch sein Gegenteil unmöglich gemacht wurden. Also müsste man, um das Verfahren objektiv zu beurteilen, auch in einem gewissen Anteil der Fälle bewusst falsch entscheiden. Das ist aber

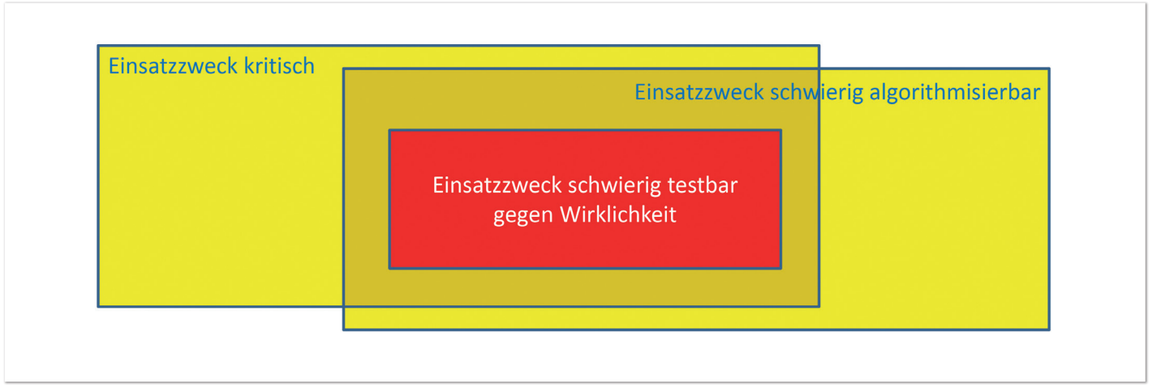


*Abbildung 5: Stufen der Gefährlichkeit des Einsatzzwecks (© Thomas Matzner)*

moralisch hochproblematisch, wie in dem Abschnitt über Konsequenzen fehlerhafter Kreditvergabe oder Freilassung gefährlicher Personen gezeigt.

## Angst vor KI ist fehlgeleitet

Eine Vielzahl der Formate, die moralische Fragen im Zusammenhang mit IT behandeln, behandeln als Gipfel der Problematik die künstliche Intelligenz (KI). Ganze Institute gibt es für Erforschung der ethischen Konsequenzen aus KI. Wie komme ich dazu, die Konzentration auf KI in diesem Zusammenhang als fehlgeleitet zu betrachten?

Betrachten wir den Einsatzzweck eines Verfahrens, in dem KI eingesetzt wird. Ist er unkritisch, haben wir kein Problem. Das trifft etwa für KI-gestütztes Schachspiel, aber auch für außerspielerische Einsätze, etwa Predictive Maintenance zu, mit der vorausbestimmt wird, wann bestimmte technische Einrichtungen instandgesetzt werden sollen, um ihrem Ausfall zuvorzukommen. Diese ist zwar nicht vollkommen unkritisch, aber immerhin geht es bei ihr nicht um menschliche Schicksale, wodurch unsere Angst sich in Grenzen hält.

Ist der Einsatzzweck leicht algorithmisierbar, haben wir es mit Regeln zu tun, die zuvor klar bekannt sind und über eine Spezifikation in das Informatiksystem überführt werden. Hier gibt es keinen Grund für den Einsatz von KI, weshalb sie auch etwa bei naturwissenschaftlich-technischen Berechnungen oder Bestellabwicklungssystemen keine Rolle spielt.

Ist der Einsatzzweck schwierig algorithmisierbar und leicht gegen die Wirklichkeit zu testen, können wir jeden Lösungsweg, auch solche auf Basis von KI, durch hinreichende Tests gegen die Wirklichkeit absichern. Hierzu gehört etwa die Gesichtserkennung. Jeder Betreiber, sei es eine Privatperson mit ihrer Fotosammlung oder der Geheimdienst mit der Fahndungsdatei, kann sich vergewissern, dass das System keine oder hinreichend wenig Fehler macht, ohne dabei überhaupt wissen zu müssen, ob KI darin steckt oder nicht.

Erst bei kritischem, schwierig algorithmisierbarem und schwierig gegen die Wirklichkeit testbarem Einsatzzweck wird es mulmig. Hier können wir bei keinem Lösungsweg sicher sein, ob er hinreichend verlässlich ist, ob er schon heute oder, wie oben dargestellt, in der Zukunft anfängt, zu viele falsche Resultate zu produzieren. Gäbe es einen Lösungsweg, etwa außerhalb der KI, zu dem man hinreichendes Vertrauen hat, könnte man ihn als Grundlage für einen Test gegen die Wirklichkeit jedes anderen Verfahrens verwenden, wodurch der ganze Einsatzzweck nicht mehr schwierig gegen die Wirklichkeit zu testen wäre. Vielmehr sind jedoch alle drei Kriterien gerade nicht abhängig vom gewählten Lösungsweg, sondern vom Einsatzzweck selbst.

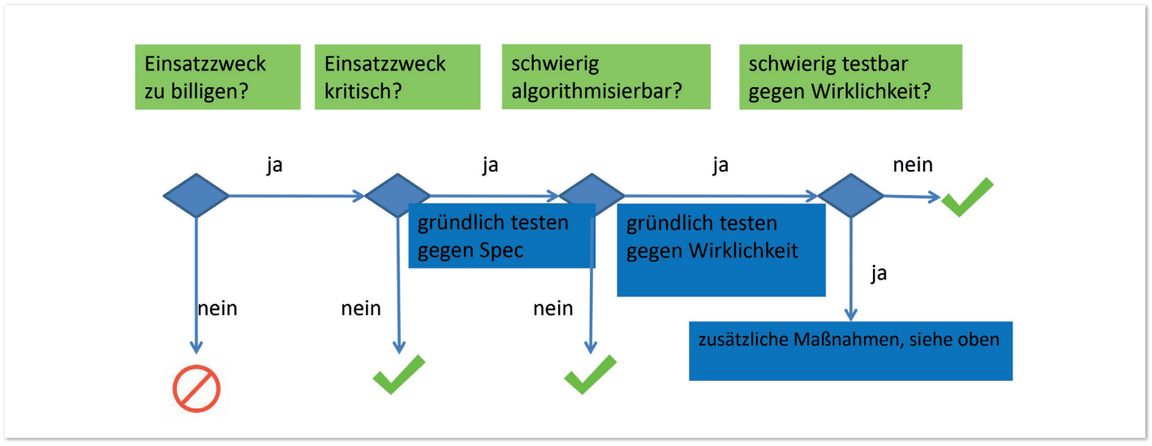


*Abbildung 6: Fahrplan für die Gestaltung moralisch integrer Verfahren (© Thomas Matzner)*

So beschreibt etwa (O'Neil 2016, 12..22) den Fall eines tatsächlich wohl fehlgeleiteten Softwareeinsatzes zur Beurteilung der Berufsleistung von Lehrern. Genaues Studium des Textes legt jedoch den Verdacht nahe, dass dort keineswegs, wie vom Hersteller der Software behauptet, KI und Big Data am Werk waren, sondern ein recht einfacher Vergleich der Schülerleistungen am Ende des vorigen und des jetzigen Schuljahrs, den jeder Laie mit dem Taschenrechner ausführen könnte. Weshalb sollten wir vor einem solchen zu einfachen Verfahren, das Schaden anrichtet, weniger Angst haben als vor einem komplizierten?

Die Angst vor KI habe ich als fehlgeleitet, nicht als unbegründet, bezeichnet. Natürlich kann man vor Fehlfunktion eines Systems berechtigte Angst haben. Allerdings entstammen alle drei Kriterien, die eine solche Angst rechtfertigen, einem Zusammentreffen dreier ungünstiger Merkmale des Einsatzzwecks, nicht den zu seiner Verfolgung eingesetzten Techniken. KI einzusetzen, ist erst bei schwierig algorithmisierbaren Einsatzzwecken sinnvoll – die Gefahr entstammt jedoch der Aufgabe und würde nicht verschwinden, wenn man etwa den KI-Einsatz verbieten würde *(siehe Abbildung 5)*.

Dazu kommt, dass bei unseren Fallstudien, wie in vielen vergleichbaren Fällen, eine Instanz über den Betroffenen entscheidet, die größer ist, als mächtiger wahrgenommen wird und aus den beschriebenen Gründen für ihn nicht transparent ist. Auch diese Umstände sind unabhängig von der eingesetzten Technik.

## Zusammenfassung: Fahrplan für die Gestaltung moralisch integrer Verfahren

Der folgende Fahrplan *(siehe Abbildung 6)* dürfte nach dem Lesen dieses Textes ohne weitere Erläuterung verständlich sein.

## Literaturverzeichnis


- Kahneman, Daniel. *Thinking, Fast and Slow.* London: Penguin, 2011.
- Matzner, Thomas. *Informatikethik.* 2. Auflage. Norderstedt: BoD, 2020.
- Nezik, Ann-Kathrin. „Wenn Maschinen kalt entscheiden." ZEIT, 10 2019: 21.
- O'Neil, Cathy. *Angriff der Algorithmen. Wie sie Wahlen manipulieren, Berufschancen zerstören und unsere Gesundheit gefährden.* Hanser, 2016.
- Wisconsin, Supreme Court of. „State vs Loomis." 13. 7 2016. *https://bit.ly/2CgErVV* (Zugriff am 9. 7 2022).
- Ziegler, Peter-Michael. „Im Namen des Algorithmus. Wenn Software Haftstrafen verhängt." c't, 12 2017: 68.



**Thomas Matzner**

*tam@tamatzner.de*


Thomas Matzner ist selbstständiger Diplom-Informatiker. Sein Hauptarbeitsgebiet ist die Konzeption von Informationssystemen. Er arbeitet in der Rolle des Requirements Engineers, Product Owners, Business Process Managers und Business Analysts.

Er unterrichtet Informatikethik im Rahmen eines von ihm aufgebauten Wahlpflichtfachs an der Technischen Hochschule Nürnberg Georg Simon Ohm.