

# Digitalgespräch Folge 78

## Synthetische Daten: Was macht sie aus und wie kommen sie im KI-Training zum Einsatz?

mit Sören Pirk von der Christian-Albrechts-Universität zu Kiel, 12. Mai 2026

<https://zevedi.de/digitalgespraech-078-soeren-pirk/>

*[Der Vorspann mit Musik und Ausschnitten aus dem Gespräch beginnt.]*

**Marlene Görger [mg]:** Herr Pirk, Sie sind Professor für Informatik an der Christian-Albrechts-Universität zu Kiel. Zu Ihren Schwerpunkten gehört auch generative KI in Verbindung mit synthetischen Daten.

**Prof. Dr. Sören Pirk [Pirk]:** Zum Glück ist es ja nicht so oft, dass jetzt ein Auto zum Beispiel mit einem Fußgänger kollidiert. Aber ein KI-Modell, das halt das Auto steuern soll in einer Gefahrensituation, muss natürlich diese seltenen Events gerade besonders gut interpretieren können. Und hier ist es so, dass wir halt diese seltenen Events besonders gut dann mit synthetischen Daten erzeugen können.

**Petra Gehring [pgg]:** Wie offen sind denn gute synthetische Daten? In der Wissenschaft spricht man ja viel von Open Data. Gerade dann, wenn es wirklich einen Unterschied macht, wie gut die synthetischen Daten sind, behalten Sie die dann zurück?

**[Pirk]:** Die meisten Anbieter fokussieren sich, weil der Prozess halt schwer ist, auf eine Kategorie von Daten. Wenn ich jetzt zum Beispiel an bildbasierte Verfahren denke, zum Beispiel nur die Kleidungsindustrie oder nur die Agrarindustrie oder so etwas, schaut euch eure Daten an. Das ist wirklich essentiell, weil wenn ich das nicht tue, dann kann ich auch kein KI-Modell trainieren. Und oftmals wird das so ein bisschen vernachlässigt.

*[Der Vorspann endet, das Gespräch beginnt.]*

**[mg]:** Über die Bedeutung des Begriffs Daten machen wir uns in der Regel wenig Gedanken. Wahrscheinlich stellen wir uns aber schon vor, dass Daten irgendwie die Welt als Quelle haben. Sie entstehen natürlich auf mannigfaltige Weise. Viele Daten werden mehr oder weniger direkt durch Messung oder Abbildung erhoben, andere vielleicht durch das Ausfüllen von Fragebögen. Ebenfalls sind wir daran gewöhnt, dass die Daten verarbeitet werden und dadurch wiederum Daten entstehen. In der Digitalität geschieht das alles überall und ständig. Praktisch relevant ist dabei der Unterschied von digitalen und nicht-digitalen Daten immer weniger. Interessanterweise taucht allerdings inzwischen eine doch neuartige Klasse abgeleiteter und nun stets digitaler Daten auf – die Rede ist von sogenannten synthetischen Daten. Der Begriff fällt zum Beispiel im Zusammenhang mit generativer

KI und dem Training von KI-Modellen. Es gibt ExpertInnen für synthetische Daten und sogar kommerzielle Anbieter, bei denen man sie beziehen kann. Das lässt aufhorchen. Was genau ist es, das durch diesen Begriff offenbar bezeichnet wird? Und welche Funktionen oder Positionen nehmen synthetische Daten im Umfeld von KI ein? Das ist unser Thema heute im Digitalgespräch. Mein Name ist Marlene Görger, ich bin Physikerin und Technikphilosophin und arbeite für das Zentrum verantwortungsbewusste Digitalisierung.

**[pgg]:** Und ich bin Petra Gehring, Professorin für Philosophie an der Technischen Universität Darmstadt. Gast und Experte heute im Digitalgespräch ist Prof. Dr. Sören Pirk. Er ist uns in die Videokonferenz aus Kiel zugeschaltet. Herzlich willkommen im ZEVEDI Podcast, Herr Pirk, wir freuen uns und sind sehr gespannt auf das Gespräch.

**[Pirk]:** Guten Tag zusammen, ich freue mich auch sehr auf das Gespräch.

**[mg]:** Herr Pirk, Sie sind Professor für Informatik an der Christian-Albrechts-Universität zu Kiel. Dort leiten Sie die Arbeitsgruppe Visual Computing and Artificial Intelligence. Bevor Sie 2023 als Professor nach Kiel gingen, haben Sie einige Jahre im Silicon Valley geforscht. Zu Ihren Schwerpunkten gehört auch generative KI in Verbindung mit synthetischen Daten, außerdem Simulation, Umweltmodellierung und digitale Zwillinge. Wir haben heute die Gelegenheit, uns von Ihnen ganz genau erklären zu lassen, was sich hinter dem Begriff synthetische Daten verbirgt. Wozu man diesen Datentyp braucht und welche Rolle er bei der Entwicklung und dem Einsatz von KI spielt. Helfen Sie uns doch zu Beginn bitte mal in diesen Begriff hinein. Was kann alles gemeint sein, wenn man von synthetischen Daten spricht? Wo kommen sie vor?

**[Pirk]:** Ja, synthetische Daten kommen da vor, wo wir mit existierenden Modellen Daten erzeugen, wo wir sie nicht aufnehmen, sondern zum Beispiel durch ein mathematisches Modell tatsächlich errechnen. Ursprünglich kommt der Begriff gerade auch aus der, sagen wir mal, Computergrafik, wo wir Bilder erzeugen durch Algorithmik. Das heißt also, hier haben wir ja dann verschiedene Verfahren, um Bildsynthese zu betreiben. Und das kommt zum Beispiel in Computerspielen zum Einsatz. Das heißt, hier definieren wir ja Algorithmen, um dann tatsächlich immersive Computergrafik zu machen und Bilder zu erzeugen. Und diese Algorithmen kann man dann natürlich auch einsetzen, um Daten zu erzeugen, die dann ähnlich aussehen wie gemessene Daten. Normalerweise als Machine Learner verwenden wir ja gemessene Daten. Das heißt, ich nehme zum Beispiel mit einer Kamera Bilddaten auf. Und diese Daten verwende ich dann ja nach einem Labeling-Prozess, nach Annotationen, dann für das Trainieren von KI-Modellen. Und genauso können wir eben mit der Computergrafik dann zum Beispiel dann auch Bilder erzeugen, direkt mit den Labels dazu. Und genauso gibt es natürlich auch in anderen Bereichen, also zum Beispiel für tabellarische Daten, für Prozessdaten, auch die Möglichkeit, dann Daten durch Modelle zu erzeugen, anstatt sie zu messen, wie sie wirklich aufgenommen werden.

**[pgg]:** Das heißt, die synthetischen Daten sind nicht einfach nur ganz fiktiv, sondern sie werden synthetisiert, um so wie eigentlich nicht-synthetische, also authentische Daten, nenne ich es jetzt mal, zu funktionieren. Habe ich das richtig verstanden?

**[Pirk]:** Genau, das ist richtig. Das heißt, wenn wir über synthetische Daten sprechen, dann versuchen wir meistens Modelle zu bauen, um dann diese Daten zu machen. Und diese Modelle, die basieren natürlich auf irgendwelchen Echtweltannahmen. Das heißt also, viele Bereiche der Wissenschaften modellieren ja mathematisch, physikalisch, biologisch, und diese Modelle können wir dann zugrunde legen, um synthetische Daten auch zu machen.

**[pgg]:** Es gibt eine Ähnlichkeitsbeziehung. Es soll irgendwie möglichst ähnlich einer bestimmten Welt von nicht-synthetischen Daten sein.

**[Pirk]:** Ja, genau. Also wir wollen halt gerne idealerweise halt Daten erzeugen, wo man keine Unterschiede mehr erkennen kann zwischen den echten Daten und den synthetischen Daten. Als Machine Learner – das ist vielleicht dann ein Stück weit technisch – geht es sehr oft darum, dass wir Modelle so trainieren wollen, dass sie auf einer bestimmten Datenverteilung, einer Data Distribution, funktionieren. Und hier ist es dann so, dass eigentlich zwischen den echten und den synthetischen Daten diese Data Distribution, die muss idealerweise überlappen und gleich sein. Und das ist halt oftmals noch sehr herausfordernd. Und deswegen forschen wir auch dazu, wie wir synthetische Daten erzeugen können, weil es halt nicht so eine ganz klare Frage ist, die man sofort beantworten kann.

**[mg]:** Welche Gründe kann es geben, synthetische Daten für das KI-Training einzusetzen?

**[Pirk]:** Ja, da gibt es mehrere Gründe für. Also wir können ja auch mal als Beispiel den Anwendungsfall nehmen, dass wir ein selbstfahrendes Auto irgendwie mit KI trainieren wollen. Das ist so, ich kann natürlich echte Autos mit Kameras ausstatten und die meine echte Welt aufnehmen lassen. Darüber bekomme ich Daten. Dann müsste ich, wenn ich die KI für diese Autos trainieren will, dann diese Daten zunächst erst mal labeln. Das heißt, ich definiere mir irgendeine Aufgabe, die ich mit meinem KI-Modell gelöst haben möchte. Zum Beispiel das Erkennen von einem Fußgänger, von einer Fußgängerin. Das heißt also, ich müsste in meinen echten Daten dann erst mal diese Fußgängerinnen dann tatsächlich labeln, maskieren in irgendeiner Form, und dann könnte ich mich hinsetzen und das KI-Modell trainieren. Jetzt ist es so, bei synthetischen Daten, da würden wir dann halt erst mal diesen Prozess nachahmen. Das heißt, wir würden versuchen, dann Daten zu erzeugen, wie sie von einem Auto gesehen werden. Und dann können wir natürlich verschiedene Szenarien erzeugen, in denen sich dann das Auto synthetisch dann in der Szene aufhält. Das Spannende jetzt, warum wir synthetische Daten wirklich erzeugen wollen, ist, dass wir damit besonders gut Randbereiche definieren können, also seltene Ereignisse definieren. Das heißt also, wenn ich jetzt zum Beispiel an die selbstfahrenden Autos wieder denke, oft habe ich bestimmte Situationen halt nicht so oft in meinen echten Daten. Also zum Glück ist es ja nicht so oft, dass jetzt ein Auto zum Beispiel mit einem Fußgänger kollidiert. Und das ist ein sehr seltenes Event. Wenn ich jetzt mehrere selbstfahrende Autos oder halt fahrende Autos hätte, dann würden nicht besonders oft Fußgänger mit Autos kollidieren, was ja auch gut so ist. Aber ein KI-Modell, was halt das Auto steuern soll in einer Situation, in einer Gefahrensituation, muss natürlich diese seltenen Events

gerade besonders gut interpretieren können. Und hier ist es so, dass wir halt diese seltenen Events besonders gut dann mit synthetischen Daten erzeugen können. Und das ist eigentlich das, worauf wir dann abzielen.

**[pgg]:** Wo ist denn dann die Abgrenzung zur Simulation? Also ein Stück weit stelle ich mir jetzt vor, da wird ja dann der Straßenverkehr simuliert, um dann die ganzen Daten dieses simulierten Straßenverkehrs auch wiederum zu nutzen. Kann man das irgendwie unterscheiden nochmal, das Erzeugen von synthetischen Daten und das Simulieren?

**[Pirk]:** Ja, das können wir unterscheiden. Oft ist es so, dass bisher synthetische Daten, gerade für bildbasierte Verfahren – das müssen wir vielleicht noch unterscheiden. Also wir reden im Moment ja über bildbasierte Verfahren entlang dieses Beispiels. Für bildbasierte Verfahren ist es momentan so, dass gerade für statische Bilder oft synthetische Daten zum Einsatz kommen. Das heißt also, ich erzeuge irgendwie eine Konfiguration, zum Beispiel dann von dem Fußgänger hin zu dem Auto, zum Beispiel kurz vor einer Kollision. Also in einer bestimmte Pose des Fußgängers gegen eine bestimmte Situation des Autos. Sie haben recht, also wir sprechen hier auch in diesem Kontext gerne von Simulationen, weil wir natürlich, wenn wir eine Simulation hätten und sich die Dinge auch über die Zeit plausibel verändern und verhalten, dass wir diese Simulation natürlich besonders gut auch für das Erzeugen von synthetischen Daten dann verwenden könnten. Das Hauptproblem dabei ist, dass diese Simulation zu bauen recht aufwendig ist. Das machen wir als Computergrafikforscher, das heißt, hier setzen wir uns hin und definieren wirklich die mathematischen Modelle, um diese Simulationen dann zu machen, wo denn Fußgänger auf der Straße herumlaufen, wo Autos herumfahren. Und das sind wirklich sehr spannende und viele Forschungsfragen, die wir hier noch zu klären haben, weil das oftmals dann wirklich die Herausforderung ist.

**[pgg]:** Das heißt, wenn es um die synthetischen Daten geht, braucht man jetzt nicht ganz alles, was man für eine gute Simulation bräuchte.

**[Pirk]:** Wir können teilweise synthetische Daten erzeugen, ohne dass wir jetzt ein sehr schwergewichtiges Simulationsverfahren anwenden müssen. Manchmal kann man auch leichtgewichtiger synthetische Daten erzeugen. Das Feld ändert sich ja gerade sehr schnell und wir haben ja neben computergrafischen Verfahren für das Erzeugen von Bildern auch die Möglichkeit, über bildgenerative KI Bilder und Daten zu erzeugen. Es gibt ja Modelle wie zum Beispiel Stable Diffusion. Diese Modelle erlauben mir, durch einen Textprompt zum Beispiel ein Bild zu machen. Es gibt inzwischen auch Modelle, die gleichzeitig nicht nur das Bild erzeugen, sondern eben auch das Label für ein Bild, was ich dann für das Trainieren einer KI verwenden kann. Das heißt also, diese Modelle erzeugen eigentlich Trainingsdaten. Das Problem ist aber, dass auch diese Modelle nicht die Daten zu einer beliebigen Qualität erzeugen, sondern auch diese Daten haben bestimmte Eigenschaften, die wir analysieren müssen und für das Training dann hinterfragen müssen, ob sie die richtigen Eigenschaften sind.

**[mg]:** Vielleicht können wir für diese beiden Fälle einmal die Herausforderung besprechen. Sie sagten ja, es gibt da ganz viele spannende Fragen im Feld der

Simulation, an denen Sie da arbeiten und eben auch in dem anderen Fall die KI-erzeugten Bilder. Vielleicht könnten Sie das noch mal ein bisschen aufschlüsseln.

**[Pirk]:** Ja, sehr gerne. Also, dann fange ich mal an mit der Computergrafik. Traditionell, die Computergrafik ist ja schon auch einige Jahrzehnte alt. Hier haben wir halt viele Verfahren, die uns ermöglichen, Geometrie zu erzeugen. Das heißt, ich erzeuge mir zunächst erst mal eine Landschaft oder bestimmte Objekte, die ich gerne in meinen Bildern sehen würde. Und ich wende da dann verschiedene Rendering-Algorithmen an. Das heißt also, wir überführen die Geometrie zusammen mit Materialien, wie wir definieren, hin zu Bildern. Das Ganze nennen wir dann Rendering. Und diese Bilder sollen in den meisten Fällen auch möglichst schnell produziert werden. Also das kennen wir halt dann zum Beispiel aus Computerspielen oder auch aus verschiedenen Filmen. Also computergenerierte Filme sind auch gute Beispiele dafür. Und hier gibt es halt unzählige Verfahren und die Qualität des Renderings, die steigt auch immer mehr. Das heißt, wir können fotorealistische Bilder erzeugen, die oftmals sehr präzise und sehr gut im Vergleich zu echten Daten aussehen. Die Herausforderung hierbei ist, dass wir die Geometrie definieren müssen. Das heißt also, wenn ich jetzt ein Auto in der Szene sehen will, dann muss ich irgendwie die Geometrie von dem Auto erzeugen. In den meisten Fällen passiert das dadurch, dass ein 3D-Artist, also ein Mensch, dieses Auto definiert und wir können dann dieses Auto dann in einer Szene arrangieren und dann zum Beispiel auf eine Straße stellen, die dann auch von einem Artist modelliert worden ist. Jetzt ist es so, dass sich halt in der Computergrafik viele Forscher auch damit beschäftigen, prozedural diese Geometrie zu erzeugen. Das heißt also, wir definieren Algorithmen, um dann diese Geometrie erzeugen zu können. Das geht zum Beispiel besonders gut für Bäume, für Straßen, also so ein bisschen, wo Struktur halt schon vorhanden ist. Für einige Kategorien von Objekten ist das sehr schwer. Wenn ich mir zum Beispiel das Gesicht von einem Menschen vorstelle, dann ist es relativ schwierig, die Gesichtszüge von einem Menschen algorithmisch zu beschreiben. Das Tolle ist, wenn wir prozedural in der Lage sind, Geometrie zu beschreiben, dann können wir auch leicht Varianzen damit erzeugen. Wenn ich diese Varianzen dann habe, dann kann ich halt verschiedene Bilder dann davon machen und diese Objekte dann fotorealistisch rendern, und das ist eigentlich genau das, was ich brauche dann für meinen Trainingsprozess von einem KI-Modell. Auf der anderen Seite haben wir dann bildbasierte generative KI. Das heißt, das sind ja dann Modelle, die Bilder erzeugen können, oftmals aus zum Beispiel Textprompts, also aus Texteingaben. Das heißt also ich beschreibe dann, was ich gerne in meinem Bild sehen möchte und das KI-Modell erzeugt dann aus diesem Textprompt dann ein Bild. Also ich kann zum Beispiel sagen, ich möchte ein Bild von einem Auto erzeugen und dann erzeugt das Modell ein Bild von einem Auto. Jetzt die Vor- und Nachteile kurz. Bei der Computergrafik ist es so, dass ich hier sehr präzise sagen kann, wie mein Bild am Ende dann komponiert sein soll. Das heißt, wo das Auto genau steht, mit welchen Materialeigenschaften es in der Szene ist, all diese Dinge kann ich halt konfigurieren. Das ist aber natürlich aufwendig, die Algorithmen zu machen, um das Auto dann am Ende fotorealistisch wie ein Foto aussehen zu lassen. Das sind halt Verfahren, die halt ineinander greifen, die dann hin zu dem fotorealistischen Bild führen. Bei der bildgenerativen KI ist das ein bisschen bequemer. Ich brauche einfach nur einen

Textprompt. Hier haben wir aber noch die Situation, dass wir nicht die volle Kontrolle haben, wie das Bild wirklich komponiert sein soll. Das heißt also, wo in dem Bild das Auto ist, habe ich möglicherweise keine Kontrolle darüber. Welche Farbe das Auto hat, kann ich vielleicht über meinen Textprompt noch beschreiben, aber möglicherweise kommt trotzdem ein anderes Ergebnis raus. Da verläuft gerade die Grenze der Forschung. Also diese Modelle werden natürlich auch immer besser. Die Controllability, die Kontrollierbarkeit dieser Modelle wird immer besser, aber hier verläuft momentan die Grenze, dass diese editability oder controllability halt noch nicht vollständig da ist. Das heißt also, ich muss ein bisschen das nehmen, was mir das bildgenerative KI-Modell dann da erzeugt. Und dann kann ich eben schauen, dass ich damit dann auch Datensätze dann erzeugen kann. Das sind so die Vor- und Nachteile.

**[pgg]:** Ich stelle mir jetzt vor, dass man schon früher oder später auch noch mal echte Daten braucht, um das irgendwie wieder rückzukoppeln. Sonst müsste man ja im Prinzip doch relativ rasch in irgendeine Schleife reingeraten, dass einfach alles immer synthetisch ist und synthetisches sich auf synthetische bezieht und irgendwie, ich sage mal, die Bodenhaftung verloren geht. Wie machen Sie das, dass irgendwie so ein Reality-Check sozusagen unterwegs noch eingebaut wird in diese Verfahren?

**[Pirk]:** Ja, das ist eine ganz spannende Frage. Also das ist halt auf der einen Seite natürlich so, wenn wir jetzt mit Modellen arbeiten. Wenn ich mir ein mathematisches Modell baue, um damit dann am Ende Bilder zu machen, dann baue ich dieses mathematische Modell natürlich basierend auf irgendwelchen Annahmen. Und diese Annahme können aus einer Machine-Learning-Sicht natürlich schnell dann auch zu sogenannten Biases, zu Voreingenommenheiten, dann führen. Und das ist eine der Grenzen der Anwendbarkeit, wenn man so überlegt. Das heißt also, was auch immer ich tue, um ein mathematisches Modell zu machen, um damit dann Bilder zu erzeugen, diese Annahmen landen in den Daten. Und das ist natürlich ein Problem. Auf der anderen Seite ist es so, wenn ich mit bildgenerativer KI arbeite, sind diese Modelle auch voreingenommen in der Form, dass ich die bildgenerativen Modelle natürlich auch auf Daten trainiert habe und was auch immer diese Daten sind, dann natürlich auch durch das Modell abgebildet werden. Das ist insbesondere deswegen spannend, weil wir ja eigentlich synthetische Daten brauchen für Fälle, die wir andernfalls schlecht messen können. Das heißt, für mein selbstfahrendes Autoszenario will ich eigentlich gerne Situationen haben, zum Beispiel kurz vor Unfällen, sodass ich eine KI trainieren kann, die dann dazu führt, dass Unfälle vermieden werden. Das heißt also, ich brauche Daten, die genau diesen Moment beschreiben, wo es gerade so kurz vor einem Unfall ist sozusagen. So, und diese Daten, die sind selten. Und hier brauchen wir eigentlich synthetische Daten. Das heißt also für die seltenen Events. Als Machine Learner sprechen wir dazu auch oft dann von dem sogenannten Long Tail einer Datenverteilung. Also die seltenen Events, die wir nur sehr selten messen, die wollen wir idealerweise eigentlich gerne synthetisieren. Und hier ist es oft so, dass das halt für viele Aufgabenkategorien auch besonders spannend ist. Also es gibt zum Beispiel das Feld der Anomalie Detection. Also kann ich hier irgendwo Daten erzeugen für Anomalien? Also das kann entweder zum Beispiel im Gesundheitsbereich sein, hat jemand eine bestimmte Krankheitserscheinung. Das kann im produzierenden

Gewerbe sein. Also wenn ich zum Beispiel Lacke bewerte und da Kratzer drin sind, das sind Anomalien, und diese Anomalien zu modellieren, das ist halt gerade das Spannende auch aus so einer synthetischen Datenperspektive. Und jetzt komme ich zurück. Das Problem ist, dass ein bildgeneratives KI-Modell diese seltenen Fälle in den meisten Fällen auch nicht richtig abbilden kann, weil es eben nicht auf diesen Daten trainiert worden ist, weil es die ja nicht gibt. Das heißt also, bildgenerative KI ist ein schöner Weg, der bequemere Weg, um synthetische Daten zu erzeugen. Aber gerade für die spannenden Fälle funktioniert dieser Weg leider nicht, weil wir halt dann diese Daten da auch nicht abgebildet sind.

**[pgg]:** Das heißt, sie müssen da dann doch irgendwie gucken, dass sie für diese seltenen Situationen, gerade für die seltene Situation, auch irgendwie handfeste, authentische Materialien haben und irgendwie dieses Modellieren von so einem seltenen Moment unterstützt wird, auch durch reale Daten.

**[Pirk]:** Genau, das ist so. Das heißt also, idealerweise, wenn ich mit synthetischen Daten arbeiten will, dann habe ich zumindest echte Daten, gegen die ich testen kann. Das wäre das Beste. Wir stellen fest, dass auch das nicht immer gegeben ist, weil halt eben diese echten Daten aufzunehmen und auch dann zu labeln extrem teuer ist oder schwierig ist. Und hier muss man dann schauen, wie man mit synthetischen Daten arbeiten kann. Ein normalerer Fall ist, dass man zumindest ein kleineres Dataset dann hat an echten Daten, gegen die ich dann testen kann, wenn ich auf synthetischen Daten trainiere.

**[pgg]:** Gibt es so einen ganz neuen Blick auf die Welt. Also sie suchen dann sozusagen ganz seltene Ereignisse, von denen es trotzdem vielleicht so die kritische Menge an Mini-Daten gibt, dass man damit was machen kann, so in Richtung Synthetisierung oder Simulation oder Modellierung.

**[Pirk]:** Ja, grundsätzlich können wir natürlich alle Szenarien irgendwie beliebig beschreiben. Also das ist halt das Spannende an synthetischen Daten. Man kann sich natürlich das alles komplett ausdenken. Die Frage ist natürlich nur, wenn wir jetzt in Richtung von Produktionsprozessen gehen. Also wenn ich mein KI-Modell in die Produktion gebe irgendwo, dann will ich natürlich auch bestimmte Robustheitsgarantien haben. Und das ist noch herausfordernd, also weil uns zum Teil dann eben die Metriken fehlen, um diese Robustheit dann wirklich zu bestimmen. Und das ist halt auch ein spannender Forschungsgegenstand, da arbeiten wir auch dran.

**[pgg]:** Das heißt, sie müssten, sagen wir mal, in einem Unternehmen, wo es um irgendeine heikle Produktionskette geht oder wie auch immer, müssten sie gezielt nach dem Ausschuss, nach den Fehlern, nach der vermurksten Situation, nach dem ganz seltenen Missgeschick irgendwie suchen und idealerweise davon Daten haben.

**[Pirk]:** Idealerweise ja, aber natürlich kann man auch probieren, mit synthetischen Daten erst mal nur vorweg zu arbeiten und dann das System auch in die Produktion zu geben. Oftmals ist es ja so, dass gerade bei diesen Anomalie-Fällen momentan Menschen auch zum Einsatz kommen, die dann am Förderband stehen und dann zum Beispiel gucken, ob in einem Lebensmittel sich vielleicht da auch eine Schraube findet

oder so, die da eigentlich nicht reingehört und solche Dinge. Das wird gemacht. Also Firmen haben ja momentan auch diese Sicherheiten eingebaut in ihren Produktionsprozessen und das ist natürlich auch sehr wichtig. Aber ich könnte natürlich jetzt auch erst mal hergehen und ein Modell auf synthetischen Daten trainieren, auch wenn ich keine echten Daten habe und dann das parallel zu menschlichen Einsatzkräften auch betreiben, um dann halt so langsam dieses Vertrauen und diese Robustheit einzustellen.

**[mg]:** Es drängt sich da ja so ein bisschen das Beispiel auch Medizin auf. Sie hatten schon Gesundheitswesen erwähnt. Gerade da will man natürlich auch richtige Ergebnisse haben, wenn man auf, sag ich mal, durch synthetische Daten angereicherten Datensätzen was lernt. Wie stellt man da denn sicher? Also wenn ich mir jetzt vorstelle, ich habe Vergleichsdaten von Patientinnen zu einer bestimmten Fragestellung und ergänze die jetzt durch synthetische Daten. Also was kann ich denn erwarten zu lernen, was jetzt nicht einfach nur eine Verdopplung der Daten ist, die ich habe, das wäre ja jetzt kein echter Gewinn.

**[Pirk]:** Also, zunächst einmal ist es da so, die Frage ist, wie kommen wir hier zu synthetischen Daten? Das ist in der Medizin besonders herausfordernd, weil wir hier noch relativ wenige Modelle haben, die wir hier zum Einsatz bringen können. Ich könnte mich ja hinstellen und sagen, ich möchte gerne das Modell von einem Herzen haben. Also wie es schlägt, wie es Blut pumpt und diese Dinge. Diese Modelle zu bauen, ist natürlich extrem aufwendig. Das heißt also, ich müsste ja ein Modell schaffen, was mir dann eine Simulation von einem Herzen definiert. Und das erstreckt sich natürlich über viele Dinge. Ich habe hier dann zum Beispiel Computational Fluid Dynamics, also Strömung des Blutes mit dabei. Ich habe Soft Bodies, das Herz als Muskel ist natürlich dann schwierig zu modellieren, all diese Dinge. Das heißt also, ich müsste erst mal in diesem Fall, wenn ich über ein schlagendes Herz spreche, eine Simulation bauen, die mir dann gegen eine Parametrisierung verschiedene Herzen dann baut und simuliert. Ich rede hier extra über Parametrisierung. Das ist das, was für uns oft spannend ist. Also wie bauen wir solche Simulationen und Modelle? Wir haben ja gerne verschiedene Parameter. Also zum Beispiel kann ein Parameter sein, wie schnell fließt Blut durch das Herz oder welche Pumpkraft hat das Herz? Wie groß ist das Herz. Das sind alles mögliche Parameter, oftmals auch Parameterketten. Und das alles muss erst mal beschrieben und simuliert werden. Und das machen wir, wenn wir mathematisch modellieren. Das heißt, dann setzen wir uns hin und denken uns solche Modelle aus. Wenn wir das Ganze in Richtung der Computergrafik ziehen, definieren wir diese Modelle auch so, dass wir geometrisch damit arbeiten können, dass wir dann am Ende wieder Bilder davon machen können. Das ist halt genau das Spannende. Und dass wir halt möglichst idealerweise das auch schnell machen können, weil wir ja auch Daten generieren wollen. Das heißt also, wenn meine Simulation jetzt mehrere Tage auf einem Großrechner gerechnet werden muss, dann bringt mir das für die Datenerzeugung leider nicht so viel, weil ich am Ende gerne bis zu einer Million Datenpunkte dann haben möchte, um mein KI-Modell trainieren zu können mit der entsprechenden Varianz dazu. Das heißt also, es sind so ein paar Stellschrauben hier, die wir hier im Auge haben müssen, wenn wir wirklich mit dann mathematischen

Modellen geometrisch modellieren, um dann Bilder zu machen. Total spannend, weil wir halt wirklich hier viele Möglichkeiten haben und viele Freiheiten haben. Wirklich auch schön und sehr tief interdisziplinär arbeiten können. Aber das sind die Herausforderungen. Und gerade bei der Medizin ist das hier so, dass wir da noch sehr am Anfang stehen. Es gibt halt natürlich Modelle vom schlagenden Herzen. Natürlich Kollegen arbeiten da dran. Aber die halt hinzubringen, so dass wir halt schnell auch Daten von einer hohen fotorealistischen Qualität vielleicht auch gegen bestimmte medizinische Bildgebungsverfahren wie jetzt CT oder MRT machen können. Da fangen wir gerade erst mit an. Aber deswegen, zurück zu Ihrem Beispiel, ist es da so, dass wir hier noch relativ wenig mit synthetischen Daten arbeiten. Aber dass das natürlich gerade in diesem Bereich extrem wichtig wäre, das zu tun. Weil wir natürlich hier eigentlich die ganze Varianz abbilden wollen, die wir so beobachten, wenn wir erkrankte Menschen dann sehen. Also, ja, dass dann halt synthetische Daten, dass diese Parameter, dieses Modellieren so machen können, dass wir wirklich bedeutungsvoll dann verschiedene Gesundheitszustände dann auch beschreiben können.

**[mg]:** Und ich frage jetzt so mal ganz konkret nach, man bräuchte jetzt synthetische Bilder von Herzen, um KI-Modelle zu trainieren, die dann später auf echte Herzen schauen sollen, um da dann was zu erkennen.

**[Pirk]:** Ja, das hängt dann natürlich von dem jeweiligen Fall ab, wonach dann die Mediziner dann da schauen. Also es kann natürlich jetzt Tumore sein, Lungenscans, das können Anomalien am Herzen sein, die da erkannt werden sollen. Das ist beliebig. Mit synthetischen Daten können wir im Prinzip alles machen, solange wir diesen Modellierungsprozess im Griff haben. Und das ist leider für die Medizin noch nicht vollständig der Fall.

**[pgg]:** Ich habe jetzt rausgehört, dass es da zwei sehr unterschiedliche Ziele gibt, die natürlich beide irgendwie im Spiel sind. Vermutlich auch besonders, wenn es dann um konkrete Anwendungen geht, also dass das praktikabel sein soll auch draußen in der Welt. Einmal natürlich die Präzision und zum anderen aber auch so eine relativ leichtfüßige Anwendbarkeit. Also es darf alles nicht zu groß und nicht zu aufwendig sein, letztlich wahrscheinlich auch nicht zu teuer sein. Also es muss irgendwie da sehr Verschiedenes unter einen Hut gebracht werden. Sie sind jetzt in der Forschung. Denken Sie immer gleich beides zusammen oder fokussieren Sie erst mal einfach Präzision und sozusagen die Reduzierung aufs Handhabbare kommt später oder umgekehrt?

**[Pirk]:** Meine Arbeitsgruppe, wir haben da schon ein starkes Augenmerk drauf, dass wir beides gleichzeitig denken. Das macht uns so ein bisschen aus, also ich trage da so ein bisschen oder vertrete beide Seiten, also einmal die Computergrafik und dann eben die KI. Das ist aus meiner Sicht halt gerade sehr spannend, weil wir halt diese Fragen alle zu lösen haben. Es ist tatsächlich so, dass wenn wir mathematisch modellieren, dann ist es oftmals so, dass wir hier vorsichtig so bestimmte Trade-offs machen müssen. Also eben aufgrund dieser Idee, dass wir dann auch wirklich KI dann damit trainieren wollen. Also Kolleginnen aus anderen Wissenschaftsbereichen, die arbeiten

natürlich an sehr aufwändigen, sehr präzisen mathematischen Modellen, um bestimmte Simulationen zu tun. Also zum Beispiel in der Strömungsforschung oder dann auch in der Biologie oder so. Und oftmals ist es dann so, dass diese Modelle nicht direkt verwendet werden können, um Bilder davon zu machen, weil die halt einfach zu aufwendig sind oder weil sie nicht in Richtung von der Erzeugung von Geometrie gedacht worden sind. Das sind zum Beispiel dann statistische Modelle. Wir können diese Modelle dann nicht direkt anwenden, um Bilder davon zu machen. Also von den Phänomenen zu machen, die diese Modelle beschreiben. Das heißt, wir müssten uns dann hinsetzen und dann die Modelle abändern, vielleicht weniger präzise machen, damit sie dann auch schnell gerechnet werden können. Das ist auch insbesondere gerade in der Robotik spannend. Also mir bringt eine reale Szene nichts, wenn ich ein Bild dann über Minuten nur erzeugt bekomme. Der Roboter soll auch trainiert werden in der Welt, soll rumlaufen in der Welt. Und dafür, um das abbilden zu können, muss er viele Beispiele bekommen. Und das heißt also, ich brauche viele, viele Bilder, um den Roboter trainieren zu können. Und das geht halt nur, wenn ich schnelle, synthetische Verfahren habe, die mir dann diese Bilder erzeugen können. Was Sie angesprochen haben, und das ist halt das Wichtige eigentlich bei synthetischen Daten: das ist die Generalisierbarkeit. Und das ist auch eine große Herausforderung, wenn wir mit synthetischen Daten arbeiten. Das heißt also, wir wollen eigentlich synthetische Daten erzeugen, sodass wir ein KI-Modell damit trainieren können, sodass dieses KI-Modell auf ungesehene echte Daten hin generalisiert. Das Problem dabei ist, dass die synthetischen Daten oftmals hin zu den echten Daten eine sogenannte Domain Gap haben. Also wir haben zwei Domains, die echte und die synthetische, und hier gibt es einfach Unterschiede. Manchmal nennt man das auch, dass wir so eine Sim-to-Real-Gap haben, also eine Simulations-zu-echt-Lücke. So, und das ist ein spannendes Forschungsfeld, weil wir halt, wir bauen mathematische Modelle, mit denen können wir Bilder machen. So, diese Bilder haben jetzt Eigenschaften, weil sie ja auf den Annahmen meines mathematischen Modells beruhen und deswegen sind diese synthetischen Bilder möglicherweise ein bisschen anders als das, was ich in echt beobachte. Auch wenn ich wirklich die Daten, die ich hier erzeugen will, wirklich stark einschränke. Und diese Lücke zu beschreiben, die Lücke zu schließen, diese Syntorial Gap zu schließen, das ist wirklich spannende Forschung und das muss passieren. Wenn ich sie nicht geschlossen habe, dann komme ich zurück wieder als Machine Learner, dann habe ich zwei verschiedene Datenverteilungen und dann ist mein Modell mit einem Bias trainiert, also auf den synthetischen Daten, sodass es halt nicht generalisieren kann zu den echten Daten. Und das ist dann ein großes Problem, denn wenn ich diese Lücken nicht hinreichend genug schließe, bringt mir das nichts, dann ist mein Modell auf falschen Daten trainiert und es kann die echten Daten dann nicht vernünftig interpretieren. Das will man natürlich nicht haben.

**[pgg]:** Gibt es da Beispiele, wo es ganz besonders hapert mit diesem Sim-to-Real-Gap? Also so Beispiele, die wir uns vorstellen können.

**[Pirk]:** Ja, also für viele bildgebende Verfahren haben wir das Ganze teilweise schon ganz gut im Griff. Also wenn wir zum Beispiel so Straßenszenen uns überlegen, wir haben halt wirklich tolle Syntheseverfahren, mit denen man Daten wirklich schön

generieren kann. Ganz besonders stark ist diese Sim-to-Real-Gap, wenn man so ein bisschen den Bildraum verlässt, also und hingeht zu dynamischen Prozessen. Also wenn ihr zum Beispiel einen Menschen habt, der sich bewegt, der rennt, der vielleicht einen Ball fängt oder wenn ein Auto durch die Gegend fährt, vielleicht wenn ich auch unebenes Terrain habe und solche Sachen. Das sind alles schwierige Sachen. Also hier haben wir in der mathematischen Modellierung halt oftmals diese Prozesse noch nicht im Griff oder noch nicht schnell genug im Griff, um dann wirklich wieder große Datenmengen anzulegen. Also viele Prozesse sind dann auch sehr unerschlossen, sagen wir mal.

**[mg]:** Wie muss ich mir denn das im Training so praktisch vorstellen, sagt man, heute trainieren wir mit synthetischen Daten, morgen mit Daten, die nicht synthetisch sind, oder wir mischen die und dann gibt es irgendwie bewährte Mischungsverhältnisse 1 zu 9 oder sowas.

**[Pirk]:** Ja, da gibt es tatsächlich verschiedene Strategien dazu. Also man kann natürlich ein Basistraining machen auf synthetischen Daten und dann mit echten Daten dann über dieses Modell drüber trainieren. Also für die Machine Learner kann ich zum Beispiel erst mal so ein Basistraining machen für eine ganze Weile, dass ich meine Weights eines Modells erst mal in eine bestimmte Richtung dann trainiert habe. Und dann kann ich mit echten Daten darüber gehen und ein sogenanntes Feintuning dann machen meines Modells. Man kann aber natürlich auch mit gemischten Daten trainieren. Und oftmals ist gar nicht so sehr die Frage, was der ideale Prozess wäre. Also was ich in der Praxis oft sehe, ist, dass die Leute einfach, weil sie überhaupt gar keine echten Daten haben, dann erst mal überhaupt mit synthetischen Daten trainieren, um überhaupt erst mal weiterzukommen. Idealzustände sind irgendwo in der Mitte. Das heißt also, hier hatte ich wahrscheinlich irgendwie mindestens 20, 30 Prozent echte Daten und das ist, was ich aus meiner Erfahrung kenne, dass wir oftmals dann die besten Trainingsergebnisse kriegen, wenn wirklich mit vielen varianzreichen synthetischen Daten gearbeitet wird und dann noch echte Daten dazugegeben werden. Und dann haben wir meistens ein Modell, was die Datenverteilung dann schön beschreiben kann.

**[mg]:** Jetzt sind Sie ja in der Situation, als Forscher selbst, die synthetischen Daten auch erstellen zu können, die Sie brauchen. Es gibt aber ja auch kommerzielle Anbieter synthetischer Daten. Wenn man mal googelt, findet man da relativ schnell. Und dann gibt es wohl auch Qualitätsstandards, besonders gute Anbieter. Was bekomme ich denn oder was kann ich denn bekommen bei einem Anbieter für synthetische Daten? Und was zeichnet da dann auch die Guten oder die Seriösen aus?

**[Pirk]:** Ja, das ist eine spannende Frage. Ich bin, muss ich sagen, ein bisschen befangen hierzu, weil ich selber auch in einem Start-up drin bin, was synthetische Daten erzeugt. Also das nur als Disclaimer hier für meine nächsten Antworten. Man muss sich erst mal fragen, welche synthetischen Daten brauche ich denn, weil halt die meisten Anbieter fokussieren sich, weil der Prozess halt schwer ist, auf eine Kategorie von Daten. Oft auch dann, wenn ich jetzt zum Beispiel an bildbasierte Verfahren denke, auch oftmals dann hier auf bestimmte Kategorien, also zum Beispiel nur die Kleidungsindustrie oder

nur die Agrarindustrie oder so etwas, weil einfach das Definieren dieser mathematischen Modelle der Computergrafik dann oftmals noch sehr schwierig ist und das halt nicht beliebig dann über alle verschiedenen Domänen hinweg definiert werden kann. Worauf man natürlich achten sollte, ist, dass man eine hohe Qualität an Fotorealismus bekommt, wenn wir über Bilder sprechen oder und entsprechend für andere synthetische Daten. Also ich kann natürlich auch synthetische Daten zum Beispiel für verschiedene Wirtschaftsprozesse definieren und hier ist das natürlich eine ganz andere Kategorie und Qualität an Daten, die da eine Rolle spielt. Wenn wir bei bildbasierten Daten bleiben, ist es natürlich so, dass idealerweise diese Daten so fotorealistisch aussehen sollen, wie es geht. Und dass die Labels dann entsprechend auch gegen meine Aufgabe, für die ich dann trainieren möchte, dann auch definiert sind.

**[mg]:** Ich bin mal neugierig auf diese Wirtschaftsdaten. Sie hatten ja ganz am Anfang auch schon mal von tabellarischen Daten gesprochen. Ich stelle mir jetzt vor, dass das wahrscheinlich eine Domäne ist, wo das eine Rolle spielt. Ja, was gibt es denn da für Beispiele, die anschaulich machen, worum es da geht, was die Schwierigkeiten sind?

**[Pirk]:** Ja, also wenn ich jetzt zum Beispiel mir vorstelle, dass ich Menshendaten habe, also zum Beispiel ich habe eine Versicherung und habe jetzt Versicherungsfälle, die da zu bearbeiten sind, dann müsste ich natürlich durch meinen synthetischen Prozess irgendwie diese verschiedenen Fälle dann abbilden. Also wenn jemand jetzt zum Beispiel einen Schadensfall einreicht und eine Erstattung möchte, da sind ja Daten, die jetzt eine Versicherung zum Beispiel dann für sich bereithält. Und wenn jetzt eine Versicherung synthetische Daten erzeugen möchte, dann müssten natürlich gegen diese Daten dann Modelle gebaut werden. Das heißt also, wie oft kommen Schadensfälle vor, wie hoch ist die Schadenssumme, all diese ganzen Fragen. Also hier kann man dann halt Modelle bauen, die halt dann statistisch funktionieren. Das heißt also ich kann mir dann synthetisch Menshendaten erzeugen oder Daten von Menschen erzeugt, die halt in solche Versicherungsfälle dann zum Beispiel abbilden. Das kann man sich natürlich auch beliebig in alle anderen Richtungen dann vorstellen. Also ich kann ja auch zum Beispiel Zeitreihendaten haben. Zum Beispiel ganz banal das Wetter zum Beispiel. Wie verhält sich das Wetter zu einem gegebenen Tag im Jahr? Da habe ich natürlich auch saisonale Veränderungen, die kann ich modellieren. So kann ich auch dann verschiedene Sensordaten modellieren, die nicht bildbasiert sind. Also zum Beispiel auch Bewegungssensoren oder TidenHub von irgendwelchen Bojen. All diese Dinge kann man sich anschauen und dann halt auch wieder mathematisch dann irgendwie ausdrücken und beschreiben. Und auch dann Varianzen beschreiben durch eine bestimmte Parametrisierung. Das geht eigentlich für jeden Prozess. Wie gut, das sei dann immer mal dahingestellt.

**[pgg]:** Ich vermute jetzt mal, dass synthetische Daten auch die ganze Frage nach Datenschutz, Personenbeziehbarkeit oder dergleichen von vornherein erübrigen. Oder ist es so, dass auch beim Themenkreis synthetische Daten irgendwie so Datenschutzfragen am Horizont bleiben?

**[Pirk]:** Ja, man darf das nicht komplett beiseitewischen, weil oftmals haben ja bestimmte Modelle auch besondere Voreingenommenheiten. Das heißt, es geht ein bisschen mehr in Richtung der bildgenerativen KI. Diese KI ist ja trainiert auf einem Datensatz und dieser Datensatz kann möglicherweise schon problematisch sein aus einer Datenschutzperspektive. Das heißt also, wenn ich jetzt Stable Diffusion als bildgeneratives KI-Modell verwende, um Daten zu erzeugen, dann kann Stable Diffusion möglicherweise datenschutztechnisch problematisch sein oder ein beliebiges anderes Modell. Es ist so, wenn ich ein mathematisches Modell baue, um damit anderweitig dann Daten zu erzeugen, dann ist die Frage, welche Annahmen gehen in dieses Modell hinein und hat das möglicherweise ein Problem? Also habe ich Biases durch den einen oder den anderen Prozess, der möglicherweise dann mit einer Datenkonformität, einem Datenschutz in irgendeiner Weise im Konflikt steht? Das muss man dabei bedenken. Grundsätzlich ist es natürlich so, wenn ich jetzt synthetische Daten erzeuge, z. B. Bilder erzeuge, dass ich damit dann natürlich, wenn ich diese Annahmen richtig getroffen habe, natürlich dann privacy-konforme Daten erzeugen kann. Also ich habe ja dann keine echten Menschen oder keine echte Bewegungen oder was auch immer dann in meinen Daten und bin dadurch natürlich dann erst mal konform.

**[pgg]:** Das könnte ja dann auch ein Argument sein, einfach die synthetischen lieber zu nutzen als andere, selbst wenn ich sie hätte, weil ich tatsächlich die Einwilligung für die konkrete Nutzung nicht habe bei den echten Daten.

**[Pirk]:** Das ist auf jeden Fall richtig. Das heißt also gerade zum Beispiel im Bereich Medizin oder auch in anderen Prozessen, wo es halt irgendwie darum geht, dass Menschen durch KI in irgendeiner Weise eingeschätzt oder bewertet werden sollen oder ihnen geholfen werden soll, ist es natürlich so, dass ich also gerade in der EU hier datenschutzkonform arbeiten muss und dann sind synthetische Daten ein guter Weg, um dann die KI-Modelle zu trainieren und dann leichter dann an Daten ranzukommen für mein KI-Modelltraining.

**[mg]:** Ich würde gerne noch mal auf das Beispiel mit der Versicherung zurückkommen. Wenn ich mir jetzt vorstelle, klar, synthetisch generiertes Bildmaterial, da kann ich einfach mit meiner Intuition drauf gucken und sagen, das sieht nicht realistisch aus, oder an dem Bild stimmt irgendwie was nicht, das ist nicht gut gelungen, wie auch immer. Wenn ich jetzt einen Datensatz über Bevölkerungsgruppen erstelle, wo Eigenschaften von Personen oder Häufigkeiten von Versicherungsfällen drin vorkommen, und ich lasse mir die gerade deswegen generieren, weil mir die echte Basis fehlt, ist es ja jetzt wahrscheinlich deutlich schwieriger, einfach nur durch draufgucken zu sagen, ist das plausibel, hilft mir das? Also ich könnte natürlich dann alles Mögliche darauf trainieren, aber kommt da dann nicht Unsinn raus? Also wie stelle ich sicher, dass mir das überhaupt weiterhilft mit der Frage, die ich habe.

**[Pirk]:** Ja, es geht natürlich immer darum, die Daten auszuwerten. Das heißt also, auch wenn wir mit synthetischen Daten arbeiten, müssen wir uns immer fragen, welche Eigenschaften haben diese Daten? Wenn man als Machine-Learner arbeitet, was ich auch gerne meinen Studenten in den Vorlesungen erzähle, ist, schaut euch eure Daten

an. Das heißt, dass es das Wichtigste ist, wird oft auch unterschätzt, also wie wichtig es ist, wirklich die Daten erst mal anzuschauen oder durch Algorithmen auszuwerten. Das ist wirklich essentiell, weil wenn ich das nicht tue, dann kann ich auch kein KI-Modell trainieren. Und oftmals wird das so ein bisschen vernachlässigt. Das ist zumindest das, was ich so in der Praxis oder dann auch bei meinen Studierenden feststelle. Das ist richtig wichtig. Und die Daten müssen natürlich die Eigenschaften haben, die man erwartet, weil sonst erzeuge ich Biases und dann sind diese Biases dann sofort auch in meinen trainierten Modellen mit drin. Das ist wichtig. Diese Analyse zu betreiben, ist nicht immer leicht. Also wenn ich jetzt zum Beispiel ein bildgeneratives KI-Modell habe und das generiert mir das Bild von einem Menschen, wie stelle ich denn dann sicher, dass das wirklich ein Bild von Menschen ist? Das ist keine leichte Frage, dazu forschen wir auch. Also zu dem Thema Metriken, wie können wir bewerten, ob bildgeneratives KI-Material da wirklich die Eigenschaften hat, nach denen wir suchen. Das ist aber eine schwierige Frage, weil ich natürlich, wenn ich irgendwas generiere mit einer KI, habe ich ja keine Ground Truth, also nichts, auf was ich mich verlassen kann. Das heißt also, ich muss irgendwie statistisch bewerten, welche Eigenschaften dann zum Beispiel ein Bild hat oder ein Datensatz hat. Das ist herausfordernd und da arbeiten wir natürlich auch dran.

**[mg]:** Ist das dann auch etwas, was man transparent macht bei fertig trainierten Modellen? Da sind synthetische Daten eingeflossen zu dem und dem Anteil oder wie auch immer? Oder ist das sowieso überall der Fall und man redet da gar nicht weiter drüber?

**[Pirk]:** Ja, dazu muss ich ehrlich sagen, dazu ist mir nichts bekannt. Also es gibt natürlich keine Gesetze oder so, die das vorschreiben momentan, ob und wie mit synthetischen Daten zum Beispiel trainiert wird. Zumindest nach meinem Kenntnisstand. Es macht natürlich Sinn, über so etwas nachzudenken. Also aber ich glaube, bisher haben wir noch nicht die Breite an Verwendung von synthetischen Daten. Ich glaube, im Moment experimentieren viele sowohl auf Industrieseite als auch in der Forschung mit synthetischen Daten. Wir erleben hier, dass die Erzeugung schwierig ist, weil wir eben entweder mathematische Modelle brauchen oder wenn wir bildgenerativ arbeiten, diese Modelle dann bestimmte Daten noch nicht erzeugen können, so wie wir sie eigentlich bräuchten. Das sind Herausforderungen, die wir versuchen anzugehen. Aber natürlich macht es auch Sinn, über solche Fragestellungen nachzudenken. Also wie können wir dann bestimmte Regeln auch schaffen für den Einsatz dieser Daten? Ganz klar.

**[pgg]:** Wie offen sind denn gute, synthetische Daten? In der Wissenschaft spricht man ja viel von Open Data. Da gibt es auch Fachkulturen. Ich sage jetzt mal, das sind vielleicht keine synthetischen Daten, sondern überhaupt Daten, wo man sowas hat wie Referenzdatensätze oder so. Also wo ganze Communities auch ganz bewusst, um Vergleichbarkeit herzustellen, auf demselben Datenpool sozusagen arbeiten und den auch gemeinsam qualitätssichern. Wie ist das in Ihrer Community? Also gerade dann, wenn es wirklich einen Unterschied macht, wie gut die synthetischen Daten sind, behalten Sie die dann zurück. Und das ist quasi eigentlich dann Ihre spezifische wissenschaftliche Leistung und damit auch sowas wie Ihr Intellectual Property oder

Ihre eigene authentische Lösung. Oder wird das geteilt, wird das der Industrie zur Verfügung gestellt und ist damit sowas wie Gemeingut?

**[Pirk]:** Ja, also das, was ich beobachte, wir haben noch nicht ganz diese Frage tatsächlich in der Community, also die Machine-Learning-Community an sich, die ist auch gerade deswegen auch sehr stark, weil sehr offen eigentlich Daten auch geteilt werden. Es gibt halt wirklich so ikonische Datensätze wie zum Beispiel ImageNet für das Trainieren von Klassifizierungsmodellen. Das sind echte Daten, aber die hat die Community natürlich erstaunlich schön mit Datenoffenheit gearbeitet und dann natürlich auch die ganze Community vorangebracht. Für synthetische Daten sind wir da, glaube ich, noch nicht ganz angekommen. Momentan haben wir es als Community hier noch schwer, wirklich diese Daten zu erzeugen. Das heißt also wirklich, diese mathematischen Modelle zu bauen, um Bilder zu machen oder bildgenerativ diese zu erzeugen, das fängt langsam an. Wir sehen auch immer mehr Arbeiten dazu. Aber es ist immer noch schwer. Was das so schwer macht, ist, dass ich ja in dem Moment eigentlich gerne dann auch echte Daten hätte, um das Ganze auch zu validieren und zu vergleichen. Das heißt, wenn ich jetzt aber synthetische Daten und echte Daten brauche, dann habe ich als Forscher – also meiner Meinung nach, ich würde ein Projekt definieren wollen – habe ich als Forscher diese zwei Herausforderungen. Also zum einen muss ich dieses mathematische Modell bauen, um Daten zu erzeugen, und auf der anderen Seite muss ich echte Daten aufnehmen und die auch noch labeln. Da gibt es wenige Beispiele bisher. Und ich hoffe, dass das mehr wird, dass wir da halt dann auch Forschung mehr in die Richtung machen können. Grundsätzlich würden diese Daten dann hoffentlich natürlich auch geteilt werden. Ich glaube, ein bisschen schwieriger ist es bei dieser Erzeugung der mathematischen Modelle, weil das natürlich wirklich dann Algorithmen sind, die dann zum Teil sehr wertvoll sind, sage ich mal so. Und da haben wahrscheinlich dann verschiedene Interessengruppen dann einfach nicht unbedingt sofort dann die Möglichkeit, dann die auch öffentlich zu machen.

**[mg]:** Ich würde gerne mal auf ein ziemlich spannendes Projekt zu sprechen kommen, das Ihre Arbeitsgruppe verfolgt. Die Wildfire Twins, also man hört schon im Namen, da geht es um Zwillinge, vermutlich digitale Zwillinge. Was ist das für ein Projekt? Können Sie mal beschreiben, was da passiert und auch ob und wie Sie da synthetische Daten einsetzen oder entstehen lassen?

**[Pirk]:** Ja, genau. Also hierzu forschen wir zu digitalen Zwillingen von Waldbränden. Da ist meine Arbeitsgruppe gefördert worden durch die EU durch einen sogenannten ERC-Consolidator-Grant. Also wir haben hier Funding bekommen, um Forschung dann in Richtung von Waldbränden zu machen. Und hier ist die Idee auch, dass wir ein mathematisches Modell bauen von Waldbränden. Das heißt also, wir modellieren geometrisch Bäume, Waldökosysteme, möglicherweise auch urbane Landschaften dann auch mit Häusern und anderen Strukturen. Um die dann tatsächlich abzubrennen und Waldbrände dann zu simulieren. Waldbrände sind natürlich ein wichtiges Thema, weil der Planet ja leider immer mehr anfängt zu brennen überall. Und wir aber noch relativ wenig eigentlich dazu verstehen, wie sich Waldbränder entwickeln und mit was für einer Intensität sich entwickeln gegen auch verschiedene

Brennstoffstrukturen. Also wenn ich jetzt einen dichteren Wald gegen einen lichtereren Wald habe, wie verändert sich der Waldbrand? Das sind alles Dinge, die wir versuchen wollen, mit diesem digitalen Zwilling dann zu untersuchen. Also digitaler Zwilling bedeutet auch, dass wir idealerweise auch rekonstruktiv arbeiten wollen. That heißt also, wir wollen Landschaften rekonstruieren und dann halt verschiedene Was-wäre-wenn-Szenarien dann idealerweise ausrechnen können. Das heißt also, dass ich für einen bestimmten Standort dann Vorhersagen treffen kann. Also wenn ich jetzt zum Beispiel an einem bestimmten Standort, einer bestimmten Stadt, eine bestimmte Brennstoffstruktur habe, kann ich irgendwas dagegen tun, um mich besser gegen einen Waldbrand aufzustellen? Das Ganze ist aus meiner Sicht besonders spannend, weil wir das eben versuchen, auch mit schnellen mathematischen Modellen zu machen und möglichst schnell dann Bilder und Daten erzeugen zu können, sodass wir KI trainieren können. Und das Ganze sollte noch einen Schritt weitergehen. Wir wollen das eigentlich gerne so machen, dass wir auch Roboter damit trainieren können. Ich nenne das gerne, wir wollen sogenannte Robot Gyms gerne bauen, also Trainings Environments, Trainingsumgebungen, in denen wir den Roboter trainieren können gegen Waldbrände, dass idealerweise dann auch autonome Agenten gegen Brände dann auch antreten können. Das ist natürlich noch ein sehr langer Weg, weil wir da noch sehr am Anfang stehen, weil erst mal das Simulieren von Waldbränden herausfordernd ist. Das Ganze dann so zu tun, dass wir auch KI damit trainieren können, ist herausfordernd. Und dann ganz am Ende, das Trainieren der Roboter ist natürlich dann auch nochmal herausfordernder.

**[pgg]:** Das erstreckt sich dann sicher nicht nur auf deutsche Wälder, sondern vermutlich auf ganz unterschiedliche Landschaftstypen rund um den Globus.

**[Pirk]:** Das ist auf jeden Fall wahr. Wir sind als Gruppe auch sehr verbunden mit verschiedenen anderen Forschungsgruppen weltweit, also sowohl nach Australien, nach Amerika, Kanada. Also das ist natürlich ein globales Problem.

**[mg]:** Eine letzte Frage zu den nicht-synthetischen Daten, da, sie haben es auch echte Daten genannt. Haben die jetzt durch diese neuen Probleme oder Anwendungsbeispiele, das sind ja nicht nur Probleme, es sind ja auch viele Möglichkeiten, die sich auftun, aber hat sich der Status echter Daten irgendwie verändert oder verschoben dadurch, dass man jetzt auch mit synthetischen arbeitet? Sind die jetzt irgendwie besonders wertvoll oder authentisch oder auch umgekehrt so ein bisschen eher old school aufwendig? Hat man Regulationen behaftet, gibt es da irgendwie ein neues Verhältnis, das man zu denen einnimmt?

**[Pirk]:** Also diese Wertigkeit, die kann ich nicht so unbedingt beschreiben. Was so oft Leute unterschätzen ist, synthetische Daten sind nicht unbedingt günstiger. Was halt spannend ist, ist, wenn ich jetzt computergrafisch oder auch mit meinem Diffusion-Modell, mit meinem bildgenerativen Modell, wenn ich damit Daten mache, dann bekomme ich die Labels umsonst. Das ist halt das Tolle. Ich muss nicht mehr einen Menschen hinsetzen, der mir zum Beispiel in einem Bild eine Maske zeichnet, in einem Objekt herum. Und das ist ja, für echte Daten ist das sehr teuer. Also, zum Beispiel, wenn ich mir medizinische Daten anschau und ich bitte einen Doktor, einen Arzt, in

einem CT-Scan einen Tumor zu labeln, das ist teuer, weil der Arzt natürlich seine Expertise dahergeben muss, um dieses Bild zu labeln, zu annotieren. Das kann ich nicht beliebig machen. Ich kann nicht Millionen von Bildern für meine KI vorbereiten. Das ist ein extrem teures Unterfangen. Wenn ich das Ganze synthetisch machen würde, wäre das umsonst. Aber das Teure in diesem Fall ist wirklich das Bauen der mathematischen Modelle und der Parametrisierung. Und deswegen sind oftmals synthetische Daten nicht unbedingt günstiger. Aber wenn ich das mathematische Modell habe, dann kann ich natürlich meinen Datenerzeugungsprozess wirklich beliebig skalieren. Das heißt also, wenn ich einmal das mathematische Modell hab, um einen Waldbrand zu modellieren oder um ein Herz zu simulieren, dann habe ich das und dann kann ich da natürlich dann beliebig Daten von erzeugen und dann halt auch wirklich sehr starke KI-Modelle trainieren. Das ist halt die Vision, die wir hier oftmals haben, dass das halt toll ist, dass Daten dann also wirklich beliebig zur Verfügung stehen.

**[mg]:** Das heißt, so Stück für Stück und Beispiel für Beispiel erschließt man so die wichtigsten Situationen unserer Welt und versucht, die zu modellieren, um die dann auch dauerhaft als Quelle für synthetische Daten zur Verfügung zu haben.

**[pgg]:** Und die, nennen wir sie mal, echten Daten werden dann vielleicht unter dem Gesichtspunkt erhoben und ausgewählt, die synthetische Welt noch zusätzlich zu informieren. Also haben dann vielleicht mehr so einen assistierenden Charakter.

**[Pirk]:** Ja, das ist eine interessante Perspektive, das kann man auf jeden Fall so sehen. Wenn ich wirklich den Syntheseprozess wirklich vollständig verstanden habe und ich damit eine Datenverteilung wirklich erzeugen kann, die ähnlich der ist, die ich in echt messen kann, dann ist das tatsächlich so. Also ist halt so die Idee auch, dass ich ja was halt schön wäre, ist, wenn ich ein seltenes Event vielleicht zum Beispiel nur einmal bräuchte und dann halt beliebig oft synthetisieren könnte. Das ist natürlich spannend, aber das ist noch schwierig oftmals.

**[pgg]:** Wir haben uns gefragt, was in Ihrem Fachjargon der Name für die nicht-synthetischen Daten ist. Also echte Daten oder empirische Daten oder richtige Daten oder sagt man irgendwas?

**[Pirk]:** Ja, ich glaube, wir brauchen tatsächlich mal einen guten Namen. Es ist so, dass wir oft von echten Daten reden, obwohl das nicht ganz synthetische Daten sind auch echt. Also oft, was ich in der Community höre, ist, dass über synthetische Daten reden, dass Leute das Fake-Daten nennen. Das gefällt mir persönlich nicht so, weil die sind ja nicht gefaked oder nicht falsch. Aber sie sind eben synthetisch. Jetzt andersrum für echte Daten habe ich leider auch im Moment keinen besseren Begriff.

**[mg]:** Und damit ist dieses Digitalgespräch zu Ende und wir bedanken uns bei Sören Pirk von der Christian-Albrechts-Universität für dieses spannende Gespräch und die interessanten Einblicke. Viele Grüße nach Kiel. Und wie immer auch vielen Dank an Sie, liebe Zuhörerinnen und Zuhörer, für das Interesse und die Aufmerksamkeit. Wenn Sie mögen, hören wir uns wieder in drei Wochen zur nächsten Folge des Digitalgesprächs, einem Podcast von ZEVEDI, dem Zentrum Verantwortungsbewusste Digitalisierung.

*[Der Abspann mit Musik und Ausschnitten aus dem Gespräch endet.]*



This work is licensed under CC BY-NC-ND 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>