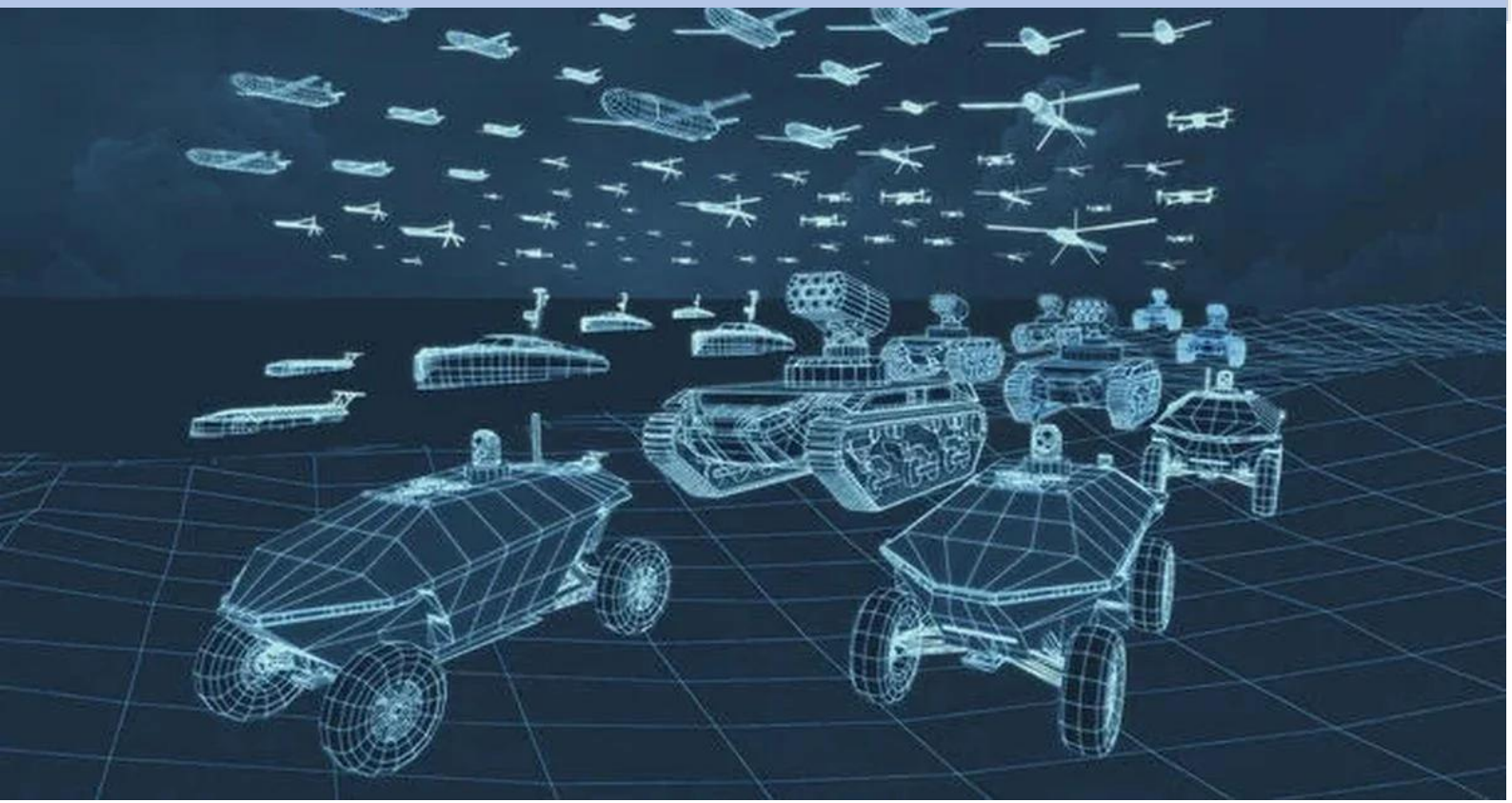




**Sind autonome
Waffensysteme objektiver als
menschliche Soldat(inn)en?**



Einstieg zu autonomen Waffensystemen (AWS)

Aufgabe 1)

Schaue dir folgende Videos in dieser Reihenfolge an.



[Merry Christmas Buddy](#)
[Scene aus Iron Man 3](#)



Autonome Waffensysteme (AWS) sind smarte KI-Drohnen, die ohne menschliche Kontrolle Ziele erkennen und ausschalten können. Auf dem Titelblatt siehst du, welche Formen Drohnen alles haben können.

Interpretiere, wie Autonome Waffensysteme (AWS) in der Iron Man Szene dargestellt werden.



[Sci-Fi Short Film](#)
[“Slaughterbots” | DUST](#)



Interpretiere, wie Autonome Waffensysteme (AWS) im Kurzfilm Slaughterbots dargestellt werden.

Aufgabe 3)

(a) Vergleiche, wie Autonome Waffensysteme (AWS) in den beiden Videos dargestellt werden.

(b) Überprüfe, ob du in (a) auf folgende Kriterien eingegangen bist:

- Nützlichkeit fürs Allgemeinwohl
- beabsichtigte Wirkung der Videos

Ergänze ggf. deine Antwort.

Aufgabe 4)

„Der Prozessor kann 100x schneller als ein Mensch reagieren. Die [zitterartige] Bewegung dient als Anti-Sniper Funktion. [...] Früher sagte man, nicht Waffen töten Menschen, sondern Menschen. Nun, die Menschen tun es nicht. Sie werden emotional, verweigern Befehle und setzen sich erhabene Ziele.“ (Slaughterbots: 1:36 – 1:45)

Wie denkst du aktuell über AWS, sind sie objektiver als menschliche Soldat(inn)en? Nimm Stellung und führe mindestens 2 Argumente für deine Position an.

Dudeneintrag zu **,objektiv‘**:

1. tatsächlich (Bsp. objektive Tatsachen);
2. nicht von Gefühlen, Vorurteilen bestimmt; sachlich, unvoreingenommen, unparteiisch

Im Laufe des Hefts wirst du dich näher damit beschäftigen, wie die KI hinter dem AWS eigentlich funktioniert und was es beim Design eines AWS zu beachten gilt. Dafür wirst du in den nächsten beiden Abschnitten *einfache KI-Modelle selbst erstellen und trainieren*. Dabei lernst du technische Grundlagen, die nötig sind, damit ein AWS *Freund und Feind voneinander unterscheiden* kann. Im folgenden Abschnitt beschäftigst du dich erstmal mit dem guten *Auswählen von Daten*, die für das Training benutzt werden. Diese werden ‚Trainingsdaten‘ genannt.

Was sind gute Trainingsdaten?

Aufgabe 1)

- ☐ Öffne auf dem Laptop, dem Tablet oder zur Not auf dem Handy die Webanwendung [Teachable Machine](#).
- ☐ Erstelle 2-3 Klassen. Diese könnten beispielsweise “Smartphone”, “Stuhl” und “Mensch” heißen. Wichtig ist, dass es Objekte im Raum gibt, die in deine Klasse fallen, damit du dein Modell später daran testen kannst.
- ☐ Benutze lizenzfreie Bilder. Diese kannst du selbst aufnehmen oder z. B. bei [Pixabay](#) finden und herunterladen. Finde für jede Klasse mindestens 5 Bilder und lade sie in der entsprechenden Klasse hoch.



Hinweis: Bilder mit der Dateiendung .webp können nicht geladen werden.
Verwende daher andere Bildformate!

Aufgabe 2)

- a) Trainiere nun dein eigenes KI-Modell, indem du den Button „Modell trainieren“ anklickst. Teste nun mithilfe von verschiedenen Objekten im Raum, die deinen Klassen entsprechen, wie gut dein Modell funktioniert. Benutze dafür deine Kamera.
- b) Klicke anschließend auf den Button “Erweitert”. Hier findest du die Parameter: “Epochen” und “Batch-Size”. Variiere diese beiden Parameter und probiere aus, welche Werte die besten Ergebnisse erzielen.
- c) Erkläre kurz, wozu die beiden Parameter dienen.

Aufgabe 3)

- a) Versuche nun, das von dir trainierte KI-Modell zu täuschen. Finde mindestens 4 Strategien, mit denen man das KI-Modell erfolgreich täuschen kann, wodurch es immer wieder falsche Vorhersagen trifft.

Hinweis: Falls du nicht weiterkommst, hole dir Hilfekarten von der Lehrkraft.

- b) Beschreibe stichpunktartig mindestens 4 Strategien, um dein KI-Modell zu täuschen.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____

- c) Entwerfe anhand deiner Erkenntnisse aus e) nun 4 - 5 Kriterien für gute Trainingsdaten, die verhindern, dass diese Täuschungsstrategien funktionieren.

1. _____
2. _____
3. _____
4. _____
5. _____

- d) Nenne einen anderen Zweck, den Kriterien für gute Trainingsdaten neben dem Verhindern vom aktiven Täuschungsversuchen noch haben.

Gute Trainingsdaten => Repräsentative Trainingsdaten

Kriterien für gute Trainingsdaten sorgen dafür, dass das KI-Modell beim Einsatz in der echten komplexen Welt noch funktioniert und Objekte auch unter ungünstigen Umständen immer noch der richtigen Klasse zuordnen kann. Kriterien für gute Trainingsdaten erfüllen damit das Ziel, dass die Trainingsdaten die echte Welt gut *repräsentieren* (abdecken).

Aufgabe 4)

Erläutere, welche Konsequenzen es für den Einsatz von Autonomen Waffensystemen haben könnte, wenn die Trainingsdaten nicht repräsentativ genug wären. Beziehe dich auf folgende Szenarien.

- a) Die KI des AWS wurde nur mit militärischen Daten trainiert, also mit militärischen Radaranlagen, Panzern und Soldaten. Diese kann es voneinander unterscheiden.

- b) Die KI des AWS wurde zu 90% mit militärischen Daten und zu 10% mit zivilen Daten trainiert. Es kommen also ein paar zivile Mobilfunkantennen, Autos und Zivilpersonen vor. Jedoch gibt es in den Trainingsdaten keine Antennen auf zivilen Gebäuden, sondern nur bei militärischen Radaranlagen.

- c) Die KI des AWS wurde zu 50% mit zivilen und zu 50% mit militärischen Daten trainiert. Jedoch sind die militärischen Daten etwas eingeschränkter, weil man von den anderen Staaten keine Fotos ihrer neuen Militärtechnologien hat. Deswegen muss man die KI hauptsächlich mit alten Panzern, Flugzeugen usw. und verpixelten Satellitenüberwachungsbildern aus großer Distanz trainieren.

Das Ziel vom Training ist es, dass das KI-Modell am Ende nicht nur auf den Trainingsdaten, sondern auch auf den Testdaten möglichst häufig die richtigen Vorhersagen und möglichst selten falsche Vorhersagen trifft. Die einfachste Möglichkeit die Qualität eines KI-Modells anzugeben ist über die Accuracy. Sie gibt die Genauigkeit eines KI-Modells an und ergibt sich so:

$$\text{Accuracy} = \frac{\text{korrekte Vorhersagen}}{\text{alle Vorhersagen}}$$

Aufgabe 5)

Genauigkeit pro Epoche

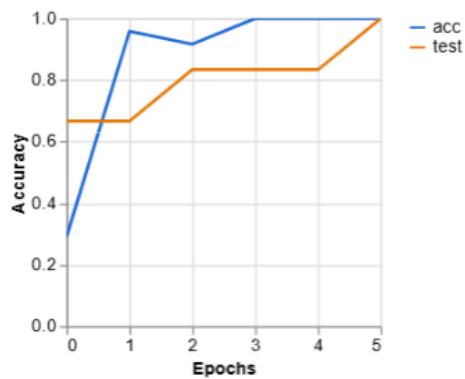


Abbildung 1: Diagramm A

Genauigkeit pro Epoche

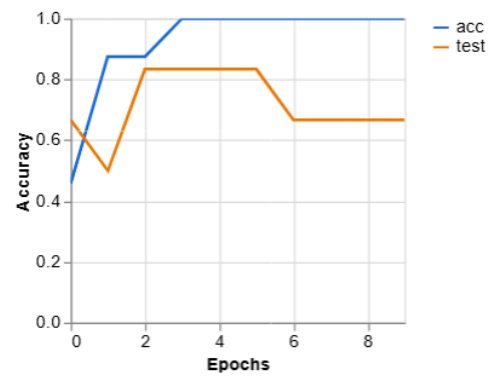


Abbildung 2: Diagramm B

acc: Accuracy auf den Trainingsdaten

test: Accuracy auf den Testdaten

- a) Begründe den Verlauf von acc und test im Diagramm A.

- b) Stelle eine Hypothese auf, warum acc und test im Diagramm B anders verlaufen.

- c) Recherchiere den Begriff "Overfitting" im Kontext von KI-Bilderkennung und überprüfe damit deine Hypothese. Notiere die richtige Antwort, falls du eine andere Hypothese hattest.

- d) Nenne anhand des Diagramms eine Möglichkeit Overfitting entgegenzuwirken.

Hinweis: Falls du nicht weiterkommst, hole dir Hilfekarten von der Lehrkraft.

- e) Erläutere, welche Konsequenzen Overfitting bei Autonomen Waffensystemen haben könnte.

Du hast dich nun mit damit beschäftigt, was repräsentative Trainingsdaten sind und warum man sie für KI-Modelle, insbesondere bei AWS braucht. Im nächsten Abschnitt wird es darum gehen, wie man ein KI-Modell so implementieren kann, dass es möglichst effektiv wird. Denn das müssen AWS sein, um nur feindlichen Soldat(inn)en, aber möglichst keinen Zivilpersonen zu schaden. Eine Bedingung für Effektivität sind repräsentative Trainingsdaten und die Vermeidung von Overfitting. Was sonst noch eine Rolle spielt, entdeckst du im nächsten Abschnitt.

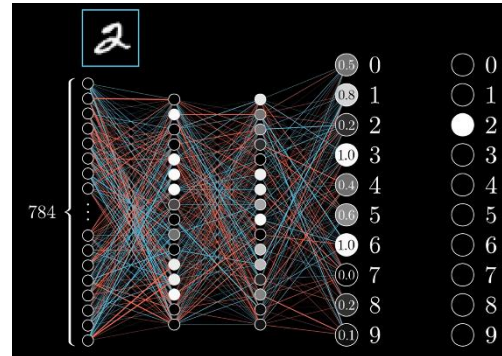
Effektivität – Von welchen Faktoren hängt die Effektivität von Künstlichen Neuronalen Netzen (KNN's) ab?

Aufgabe 1

Du wirst dir gleich ein Video ansehen. Sieh dir vorher folgende Aufgaben an, damit du sie parallel stichpunktartig bearbeiten kannst.

Beschreibe, ...

- a) ... was der Input für das neuronale Netz ist und wie Neuronen damit zusammenhängen.



- b) ... was der Output des neuronalen Netzes ist und wie Klassen damit zusammenhängen.

- c) ... wozu die Hidden Layer grob gesagt dienen.

- d) ... wozu der Bias dient.

- e) ... wozu die Gewichte dienen.

- f) ... wozu die Aktivierungsfunktion (hier die Sigmoid-Funktion) dient?

Sieh dir das YouTube Video „[Aber was ist ein neuronales Netz?](#)“ vom Beginn bis zum Zeitpunkt **12:42** an. Mache dir währenddessen Notizen zu den Fragen.



Aufgabe 2

Nun sollst du selbst ein einfaches Künstliches Neuronales Netz (KNN) erstellen und trainieren. Der Datensatz besteht aus Bildern mit 2x2 Pixeln. Diese Bilder *sollen* vom Netz entweder als *dunkel*, *grau* oder *hell* klassifiziert werden.

klassifizieren: einem Bild wird eine Klasse zugeordnet

- ☐ Öffne die Webanwendung [NetVisard](#).
- ☐ Wähle den Datensatz Graustufen 2x2 aus,
- ☐ lade ihn herunter,
- ☐ entpacke die Zip-Datei
- ☐ und lade ihn auf NetVisard hoch.



- a) Erstelle ein neues neuronales Netz, indem du auf „neues Netzwerk erstellen“ klickst. Jetzt erscheint ein Fenster, indem du die Anzahl der Neuronen pro Neuronenschicht angeben kannst. Jede Zahl, die du angibst, steht für eine Schicht. Überlege dir, wie viele Neuronen du für die Ein- und Ausgabeschicht und wie viele und wie große Hidden Layer das KNN braucht, um die 2x2 Pixel Bilder in dunkel, grau und hell zu klassifizieren.
- b) *Trainiere* dein Netz so lange, bis die Trainingsstatistik auf 70% kommt. Wenn nötig, variiere die *Aktivierungsfunktion*, die *Lernrate* und die Anzahl der zu trainierenden Bilder. Probiere aus, welche Auswahl die besten Ergebnisse erzielt.

Hinweis: Beachte, dass du immer ein neues Netz erstellen solltest, wenn du diese Einstellungen änderst.

- c) Teste das KI-Modell, indem du auf „Teste X Bilder“ klickst. Notiere hier den die Trainings- und Teststatistik der letzten 5 Durchläufe (in %).

Trainingsstatistik: _____

Teststatistik: _____

Aufgabe 3

Stark positive Gewichte sind besonders blau, während umgekehrt besonders stark negative Gewichte besonders rot gekennzeichnet sind. Je näher ein Gewicht positiv oder negativ an null ist, desto transparenter ist es dargestellt. Das gleiche gilt für Aktivierungen von Neuronen in Testdurchläufen.

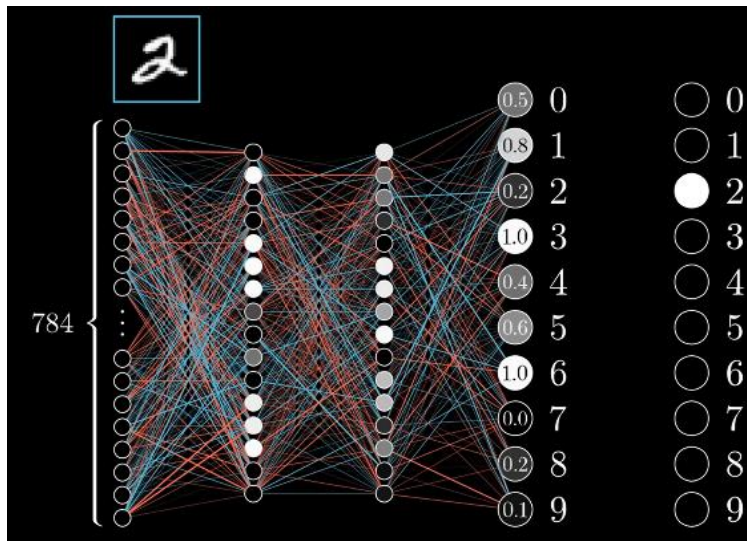
- a) Gib an, welche Neuronen in deinem Netz welchen Pixeln auf dem Bild entsprechen, und erkläre, woran man das erkennen kann.

- b) Gib an, welche Neuronen deines Netzes welcher Klasse (hell, dunkel, grau) entsprechen, und erkläre, woran man das erkennen kann.

Hinweis: Nutze ein vortrainiertes Netz mit einer guten Teststatistik, weil das Netz sonst häufiger falsche Vorhersagen treffen kann.

Aufgabe 4

Auf dem unteren Bild siehst du das Künstliches Neuronales Netz (KNN) aus dem Video. Dieses soll handgeschriebene Ziffern erkennen.



Die rechte Leiste steht für die **tatsächliche Klasse** der handgeschriebenen Ziffer. Jedes Trainings- und Testbild wird zusammen mit dieser Information gespeichert. Man nennt sie **Label**.

a) Beschreibe, was hier gerade passiert.

b) Lies dir den folgenden Infotext über Fehler durch.

Der Fehler als Berglandschaft

Dieser und folgende Info-Texte in diesem Abschnitt sind KI-generiert (bearbeitet): ChatGPT basierend auf <https://databasecamp.de/ki/backpropagation-grundlagen>

Den Fehler kann man sich wie eine riesige drei-dimensionale Berglandschaft vorstellen.

Hoch oben auf den Bergen ist der Fehler groß.

Tief unten im Tal ist der Fehler klein.

Das Ziel des Trainings ist klar:

Das Netz soll möglichst schnell ins tiefe Tal gelangen. Doch wo geht es am schnellsten bergab? Genau das beantwortet der **Gradientenabstieg**. Die genaue Definition des Begriffs ‚Gradient‘ ist kompliziert und für uns nicht so wichtig. Wichtig ist, dass der Gradient dem Netz die Richtung des **steilsten Abstiegs** zeigt — also den schnellsten Weg hinunter zu weniger Fehlern. Nach jedem Trainingsschritt kommt es einem Tal — also einer besseren Vorhersage — näher.

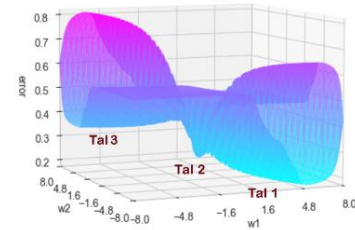


Abbildung 3: Fehlerberge und Täler

- c) Gib an, wie groß der Fehler(-berg) in unserem Beispiel für die Klasse ‚2‘ ist.

- d) Nenne Möglichkeiten, was man nach einem Trainingsschritt im KNN verändern kann, um zu einem kleineren Fehler zu kommen.

- _____
- _____

- e) Erkläre, wieso es nicht leicht herauszufinden ist, wo im KNN der Fehler liegt.

Deshalb arbeitet das KNN beim Training rückwärts

Um herauszufinden, welche Gewichte und Biases wirklich verantwortlich waren, beginnt das Netz hinten — bei der Ausgabeschicht, wo der Fehler sichtbar wird.

Von dort wird der Fehler Schritt für Schritt rückwärts durch das Netz weitergegeben:

- Welche Neuronen haben die falsche Entscheidung beeinflusst?
- Welche Verbindungen waren besonders wichtig?
- Welche wurden fast ignoriert?

➔ Welche Gewichte und Biases müssen verändert werden?

Dieses Rückwärtslaufen durch das Netz nennt man **Backpropagation** („**Fehlerrückführung**“). Dabei berechnet das Netz für jede Schicht, wie die Parameter verändert werden müssen, damit der Fehler kleiner wird.

- f) Fasse nun stichpunktartig in 4 – 5 Schritten den Trainingsprozess zusammen. Beziehe die Begriffe ‚Label‘, ‚Biases‘, ‚Gewichte‘, ‚Vorhersage‘, ‚Backpropagation‘ und ‚Gradientenabstieg‘ mit ein.

1. _____

2. _____

3. _____

4. _____

5. _____

Aufgabe 5

- a) Begründe, wie Epochenanzahl und Overfitting zusammenhängen.

- b) Unter „Netzkonfiguration“ siehst du, dass der Datensatz in Trainingsbilder und Testbilder aufgeteilt ist. Das neuronale Netz wird nur auf Trainingsbildern trainiert, während Testbilder nur zum Testen genutzt werden. *Erkläre*, warum das wichtig ist.

Hinweis: Falls du nicht weiterkommst, hole dir Hilfekarten von der Lehrkraft.

Kommen wir wieder zurück zu AWS. Hier bedeutet Effektivität des KNN insbesondere, wie gut es beim Unterscheiden zwischen zivilen und militärischen Zielen ist. Das kann z. B. bedeuten, ob das AWS ein Wohnblock von einer Militärbasis unterscheiden kann. Im letzten Abschnitt hast du bereits festgestellt, dass die Effektivität eines KNN stark von repräsentativen Trainingsdaten abhängt. In diesem Abschnitt hast du weitere technische Faktoren gelernt, von der die Effektivität eines KNN abhängt.

Aufgabe 6

Erläutere, wie die *Effektivität (test accuracy)* eines AWS mit folgenden Faktoren zusammenhängt. Benutze Beispiele, bei denen das AWS zwischen *Wohnblock* oder *Militärbasis* unterscheiden muss.

a) Anzahl der Neuronen und Hidden Layers

b) Geeignete Biases und Gewichte (durch Backpropagation und Gradientenabstieg)

c) Epochenanzahl

d) Aufteilung der Daten in separate Trainings- und Testdaten

Zusatzaufgabe

Wähle nun den Datensatz Flaggen 6x6 Rauschen aus, lade ihn herunter, entpacke die Zip-Datei und lade ihn auf NetVisard hoch. Der Datensatz besteht aus flaggenähnlichen Bildern, die mit einem Rauschen unterlegt also verpixelt sind. Diese werden vom Netz später entweder als *vertikal* (v) oder *horizontal* (h) erkannt.

- Erstelle ein neues geeignetes neuronales Netz für diesen Zweck und trainiere es so, dass es die bestmöglichen Ergebnisse erzielt.
- Erkläre, welche Auswirkungen die Lernrate auf die *Effektivität* des KNN hat und warum.

Tipp: Denke noch einmal an die Fehlerberg-Metapher.

Du kennst nun die technischen Grundlagen von AWS. Zuerst hast du dich mit den Trainingsdaten und danach mit dem Aufbau von KNN's inklusive dem Lernprozess in den hidden layers beschäftigt. Doch bisher blieb unklar, was ein KNN in den Hidden Layers eigentlich lernt. Im nächsten Abschnitt geht es um Transparenz. Hier werden wir Licht in die Hidden Layers werfen, um die Gründe für Klassifizierungen nachzuvollziehen.

Transparenz – Warum sollten die Gründe für Klassifizierungen von AWS transparent sein?

Layerwise-Relevance-Propagation (LRP) ist ein Verfahren, um die **Gründe von KI-Klassifikationen sichtbar zu machen**. Zur Veranschaulichung ein Beispiel: das linke untere Bild wird als Schloss erkannt. Mithilfe von LRP wird für jeden Pixel des Bildes berechnet, wie relevant er für die Klassifizierung als Schloss war. Diese berechnete Relevanz aller Pixel zeigt das rechte untere Bild, eine sogenannte *Saliency Map* (deutsch: *Relevanz-Karte*). Besonders relevante Pixel sind stark rot dargestellt, während Pixel, die stark gegen die Klassifizierung sprechen, stark blau dargestellt sind.

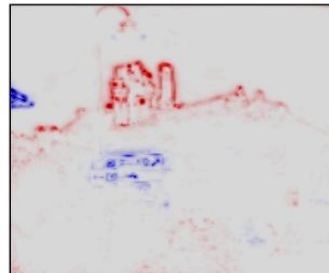
Input



Output

Schloss

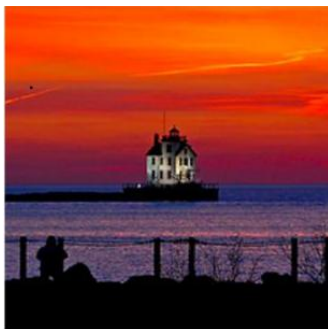
Saliency Map (mit LRP)



Aufgabe 1

Gib den Grund für die folgenden Klassifikationen an.

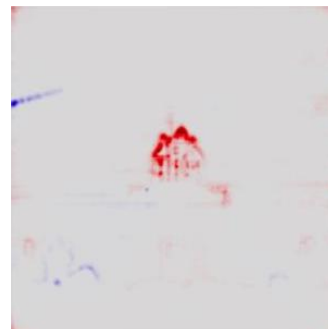
1. **Input**



Output

Haus

Saliency Map (mit LRP)



Grund: _____

2. Input



Output

Auto

Saliency Map (mit LRP)



Grund: _____

Hier ist die Klassifikation richtig und auf den Saliency Maps werden die Gründe dafür sichtbar. Bei dem Haus-Bild wird hauptsächlich das Haus rot markiert und beim Auto-Bild die Autos. Problematisch wird es aber in zwei anderen Arten von Fällen: Erstens, wenn die Klassifikation und deren Gründe falsch sind. Und zweitens, wenn die Klassifikation des KNN richtig ist, aber aus den falschen Gründen erfolgt.

Aufgabe 2

Folgende Bilder sind von einem KNN richtig klassifiziert worden, aber aus den falschen Gründen. Erkläre, welche Gründe das KNN für folgende Klassifikationen hatte und wie es dazu gekommen sein könnte.

Hinweis: Hier sind neutrale Pixel grün dargestellt.

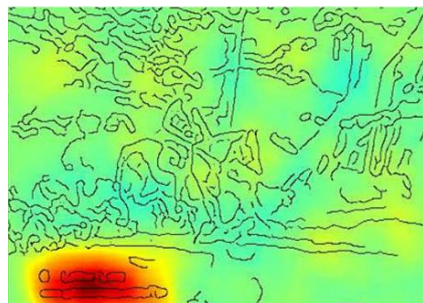
a) Input



Output

Pferd

Saliency Map (mit LRP)

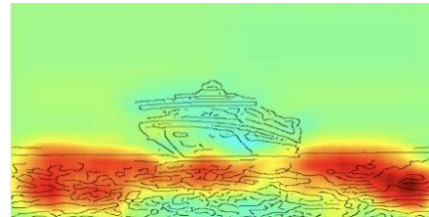


Grund:

b) **Input** **Output** **Saliency Map (mit LRP)**



Schiff



Grund:

Aufgabe 3

Sieh dir folgendes Video an: [Kluger Hans - Das Pferd, das rechnen konnte.](#)



↑
Pferd

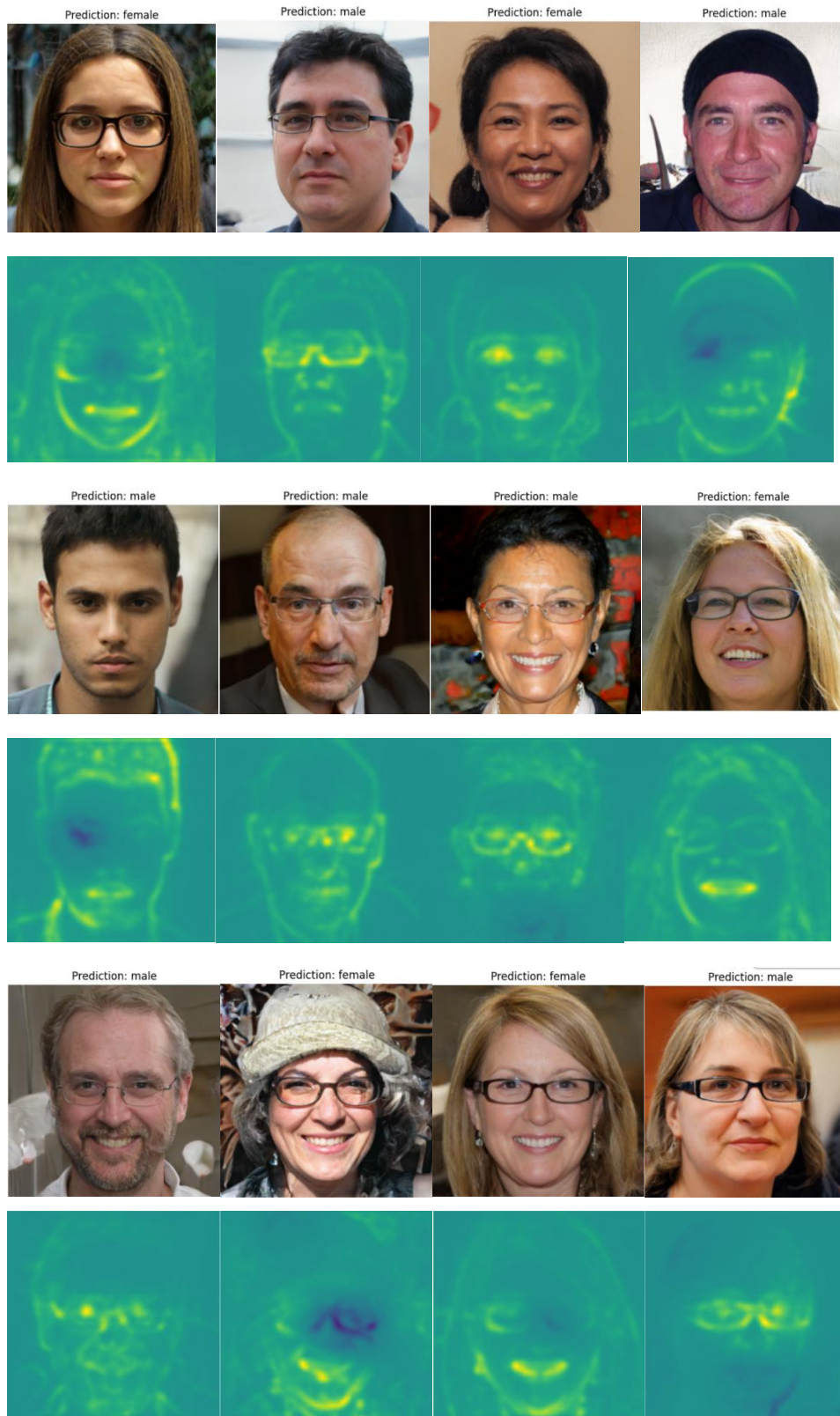
↖
Mathelehrer

a) Beschreibe, was der Trick vom Klugen Hans war, um mathematische Aufgaben zwar aus den falschen Gründen, aber meist richtig zu lösen.

b) Erkläre, was Clever Hans Tricks mit KNN's zu tun haben.

Aufgabe 4

Ein KNN wurde darauf trainiert, Männer und Frauen zu unterscheiden. Unten siehst du Testbilder mit Vorhersagen des KNN. Relevante Pixel sind gelb dargestellt, während neutrale türkis und gegenteilige Pixel blau dargestellt sind.



a) Nenne Clever Hans Tricks, die für die Klassifikation benutzt werden.

- _____
- _____
- _____
- _____

b) Erläutere, welche der Clever Hans Tricks mit Stereotypen zusammenhängen und warum sich Stereotype plötzlich in einem KI-Modell wiederfinden.

c) Begründe, wie es selbst bei komplett korrekten Labels nur durch die Auswahl der Trainingsdaten Stereotype in ein KNN schaffen können.

Aufgabe 5

Ordne das Clever Hans-Problem und Overfitting in den Lernprozess von KNN's ein. Verwende dabei die Begriffe „Gradientenabstieg“, „erstbestes Tal“, „niedrigstes Tal der Umgebung“, „Trainingsdaten“ und „Testdaten“.

Kommen wir nun wieder zurück zu AWS!

AWS-Design und AWS-Einsatz: Stellen wir uns vor, das US-Militär setzt eine autonome Drohne ein. Diese trifft eigenständig Entscheidungen, welche Ziele sie eliminiert. Ihre Test Accuracy beim Erkennen von Zivilpersonen liegt bei 99,4% und bei feindlichen Soldat(inn)en bei 99,1%. Es gibt jedoch einen menschlichen Operator, der sie die ganze Zeit per Video-Live-Stream überwacht und im Notfall stoppen könnte. Um dem menschlichen Operator das Eingreifen zu ermöglichen, informiert das AWS den Operator 5 Sekunden vor Eliminierung der Ziele über die geplante Eliminierung und die Klassen der Ziele.

Szenario: Das AWS erkennt zwei weibliche Aufständische. Tatsächlich sind die Ziele sehr weit entfernt, sodass der Operator auf dem Live-Stream wenig erkennen kann. Jedoch meint er beim Heranzoomen wage einen nach vorn gerichteten Waffenlauf zu erkennen und greift nicht ein, auch weil das AWS sonst immer richtig liegt. Tatsächlich sind es aber zwei Wanderinnen, von denen eine mit einem Wanderstock auf den nächsten Berg zeigte. Für das AWS war hingegen ausschlaggebend, dass die beiden Rucksäcke mit Tarnfarben trugen und weiblich waren, weil keine solchen Zivilpersonen im Training vorkamen.

Aufgabe 6

a) Erkläre, warum hier ein Clever Hans Problem vorliegt.

b) Begründe anhand dieses Szenarios, wieso die Gründe für Klassifizierungen von AWS transparent sein sollten.

c) *Erläutere anhand des Szenarios*, warum selbst Saliency Maps manchmal nicht ausreichen könnten, um zu verhindern, dass ein AWS Zivilpersonen anvisiert und tötet.

In diesem Abschnitt hast du die Gründe für Klassifikationen von KNN's untersucht. Dabei hast du festgestellt, dass AWS transparent sein sollten, damit Clever Hans Tricks entdeckt werden können, statt zu tragischen Fehlern zu führen, die Zivilpersonen das Leben kosten. Diese Fehler können zufällig durch schlecht gelernte Muster entstehen. Sie können aber auch durch menschliche Stereotype bei der Auswahl der Trainingsdaten entstehen. Dadurch stellt sich die Frage, ob KNN's von AWS tatsächlich neutraler sind als menschliche Soldat(inn)en.

Neutralität – Ist ein AWS neutraler als menschliche Soldat(inn)en?

neutral ≠ objektiv: eine neutrale Behauptung muss *nicht faktisch richtig* sein, *nur unvoreingenommen und von Emotionen unabhängig*; eine objektive Behauptung muss beides sein

Um die obige Frage zu beantworten, brauchen wir zunächst einen Vorschlag für ein möglichst neutrales AWS. Damit man feststellen kann, ob es überhaupt neutral ist, muss es dazu noch transparent sein.

Aufgabe 1

Entwerf einen groben Vorschlag, wie man ein AWS möglichst *transparent* und *neutral designen* oder *einsetzen* kann, um zu vermeiden, dass menschliche Stereotype in die Klassifikation miteinfließen oder es zumindest nicht zu Kriegsverbrechen kommt. Berücksichtige Probleme wie z. B. Overfitting, Clever Hans Tricks und die eingeschränkte Verfügbarkeit von militärisch repräsentativen Daten.

- _____
- _____
- _____
- _____
- _____

Um herauszufinden, ob ein AWS neutraler ist als menschliche Soldat(inn)en, brauchst du zunächst einmal Informationen über die Neutralität von menschlichen Soldat(inn)en. Deswegen hier beispielhaft die Ergebnisse einer viel zitierten Umfrage des US-Militärs:

Ergebnisse einer Umfrage vom US-Militär während des 2. Golfkriegs

Bereich	Ergebnisse
Überzeugungen	• Nur 47 % der Soldiers und 38 % sind überzeugt, dass Zivilpersonen mit Würde und Respekt behandelt werden sollten • Weit über ein Drittel der Soldaten und Marines gab an, dass Folter erlaubt sein sollte, sei es, um das Leben eines Kamerad(inn)en zu retten oder um wichtige Informationen über Aufständische zu erhalten. • 17 % der Soldaten und Marines stimmten zu oder stimmten voll und ganz zu, dass alle Zivilpersonen als Aufständische behandelt werden sollten • Etwa 10 % der Soldiers und Marines geben an, Zivilpersonen falsch behandelt zu haben (unnötige Zerstörung von Eigentum, unnötiges Schlagen und Treten von Zivilperson)
Verhalten	• 45 % der Soldiers und 60 % der Marines würden Kamerad(inn)en nicht melden, wenn diese eine unschuldige Zivilperson <i>verletzt oder getötet</i> hätten • < 50% würden ein Teammitglied wegen unethischen Verhaltens melden
Melden von Fehlverhalten	• trotz ethischer Schulung berichteten 28 % der Soldat(inn)en und 31 % der Marines, dass sie mit ethischen Situationen konfrontiert waren, in denen sie nicht wussten, wie sie reagieren sollten
Entscheidungs-fähigkeit	• hohes Aggressionsniveau, intensive Kampferfahrungen oder denen psychische Probleme verdoppelten Wahrscheinlichkeit von Falschbehandlung von Zivilpersonen
Emotionen	• neigen bei Wut eher dazu, irakischer Zivilpersonen zu misshandeln und generell unethisches Fehlverhalten doppelt so stark • Kampferfahrung, insbesondere der Verlust eines Teammitglieds, stand in Zusammenhang mit einer Zunahme von ethischen Verstößen.

Arkin, Ronald (2007): *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, Atlanta: Georgia Institute of Technology Mobile Robot Laboratory, Technical Report GIT-GVU-07-11, <https://sites.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf>, S. 7f. (KI-Übersetzung mit DeepL: bearbeitet und zusammengefasst).

Aufgabe 2

Stelle stichpunktartig dar, wie neutral oder nicht die US-Truppen im 2. Golfkrieg waren.

- _____

- _____

- _____

- _____

Im Grunde Gut

In den Jahren nach dem Zweiten Weltkrieg wurde Samuel Marshall einer der bedeutendsten Historiker¹ seiner Generation. [...] [Er fand heraus, dass] viele Soldaten, die keinen Schuss abgaben, sich gerade während der Ausbildung hervorgetan hatten. Marshall konnte außerdem keine Unterschiede zwischen frischen und erfahrenen Truppen feststellen, wenn es um die Bereitschaft zum Schießen ging. [...] Die amerikanische Armee nahm seine Arbeit ernst. [...] „Der gesunde Durchschnittsmensch“ schrieb Marshall „hat eine innere und ungewöhnlich uneingestandene Hemmung dagegen einen Mitmenschen zu töten.“ [...]

Nach dem Zweiten Weltkrieg begannen Historiker, Veteranen zu interviewen und stellten dabei fest, dass mehr als die Hälfte von ihnen noch nie jemanden getötet hatte. [...] Unter dem Eindruck dieser Erkenntnisse begannen die Historiker, auch andere Kriege aus dieser neuen Perspektive zu betrachten. Nehmen Sie die Schlacht von Gettysburg (1863) während des Amerikanischen Bürgerkriegs. Im Anschluss an die Kampfhandlungen wurden 27.574 Musketen gefunden. Aber mehr als 90 Prozent dieser Waffen waren noch geladen. Das war ziemlich eigenartig. Ein Soldat hätte nämlich 95%

¹ Hier wird nicht gegendert. Da es keine Übersetzung aus dem diesbezüglich neutralen Englischen ist, habe ich es so belassen.

mit Laden und 5% mit Schießen verbringen müssen. [...] Etwa 12.000 Musketen waren doppelt geladen, die Hälfte davon dreifach. Es wurde sogar ein Gewehr mit 23 Kugeln im Lauf gefunden. Das ergibt überhaupt keinen Sinn. Die Musketen [konnten nur] eine Kugeln abfeuern [...]. Was ging hier vor sich? [...] Das Überladen der Waffe war eine perfekte Ausrede, um nicht zu schießen. Und wenn die Waffe bereits geladen war, hat man sie einfach noch einmal geladen. Und ein weiteres Mal.

Bregman, Rutger (2023): Im Grunde Gut. Eine Neue Geschichte Der Menschheit, 12. Auflage, Hamburg: Rororo (Rowohlt Taschenbuch Verlag), S. 103 - 106.

Aufgabe 3

a) Gib die 3 Kernaussagen über Soldat(inn)en aus der zweiten Darstellung an.

- _____
- _____
- _____

b) Nenne die 3 Kern-Widersprüche zwischen der Umfrage des US-Militärs und der Darstellung im Buch „Im Grunde Gut“.

- _____
vs. _____
- _____
vs. _____
- _____
vs. _____

c) Nehmen wir an, beide Darstellungen über Soldat(inn)en basieren auf realen Geschehnissen. Begründe, wie man diese scheinbaren Widersprüche erklären kann.

Aufgabe 4

Vergleiche menschlichen Soldat(inn)en mit eines AWS nach deinem Vorschlag, indem du die zentralen Gemeinsamkeiten und Unterschiede skizzierst.

Aufgabe 5

Nimm Stellung dazu, ob ein AWS nach deinem Vorschlag neutraler ist als menschliche Soldat(inn)en. Begründe diese mit den Unterschieden zwischen menschlichen Soldat(inn)en und einem AWS nach deinem Vorschlag.

[illegible]

Situationsverständnis – Kann ein AWS eine Situation genauso gut verstehen wie menschliche Soldat(innen)?

In diesem Abschnitt lernst du das letzte Puzzleteil kennen, um zur Leitfrage dieses Arbeitshefts Stellung zu nehmen. Zur Erinnerung, sie lautet: Ist ein AWS objektiver als menschliche Soldat(inn)en. Du hast dich mit der Effektivität, der Transparenz und Neutralität von KNN's für AWS beschäftigt. Aber auf dem chaotischen Schlachtfeld braucht es mehr als nur einzelne Objekte zu erkennen. Auch die größere Situation kann einen Unterschied machen. Es bleibt die Frage, ob AWS dazu fähig sind.

Stelle dir eine Operation zur Aufstandsbekämpfung in einem Sikh-Dorf vor. [Ein AWS] hat einen Hinweis erhalten, nach dem sich gesuchte Aufständische möglicherweise in einem Wohnhaus verstecken. Der Hinweis ist falsch, doch das [AWS] zur Aufstandsbekämpfung weiß das nicht.

Betrachten wir eine realistische Interpretation des Autonomen Waffensystems. Es gibt zwei sich schnell nähernde mögliche Ziele, die beide eine Waffe tragen. Ein drittes mögliches Ziel folgt den ersten beiden und verursacht Geräusche, die mit gewalttätigem oder bedrohlichem Verhalten verbunden werden können.



Abbildung 4: Symbolbild für Interpretation der Situation durch AWS (KI-generiert)

Guarini/Bello (2011): Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters, in: Lin, Patrick (Hrsg.)/Abney, Keith (Hrsg.)/Bekey, George (Hrsg.), *Robot Ethics. The Ethical and Social Implications of Robotics*, Cambridge/London: MIT Press, S. 130. (KI-Übersetzung mit DeepL (bearbeitet))

Aufgabe 1

Nehmen wir an, die Personen wurden mit hoher Sicherheit als Aufständische erkannt. Nenne das weitere Vorgehen des AWS.

- _____

Aufgabe 2

Beschreibe, wie die Saliency Map des Sichtfelds vom AWS aussehen würde. Die Abbildung dient nur der Veranschaulichung. Beziehe dich auf den Text.

[Stelle dir nun statt einem AWS menschliche Bodentruppen vor.] In dem Haus befinden

sich drei Kinder und ihre beiden Eltern. Zwei der männlichen Kinder sind noch klein und spielen mit einem Ball. Beide tragen zudem den Sikh-Kirpan (manchmal auch als Sikh-„Dolch“ bezeichnet). Dieser ist ein religiöses Symbol und wird nicht als Waffe verwendet.



Abbildung 5: Realität (KI-generiert, nicht erkannt per KI)

Kurz bevor ein Mitglied der Aufstandsbekämpfungseinheit die Tür eintritt, kickt einer der Jungen seinen Ball in Richtung der Tür, und beide rennen ihm hinterher. Als die Streitkräfte das Haus betreten, sehen sie zwei junge Jungen auf sich zulaufen und eine schockierte Mutter schreien. Sie jagt den Jungen hinterher und schreit sie an, sich von den Männern an der Tür fernzuhalten; die Truppen wissen nicht, was sie schreit, da sie ihre Sprache nicht verstehen. Es ist durchaus möglich, dass die betreffenden Kräfte dies schnell als eine Situation wahrnehmen, in der zwei kleine Kinder spielen und eine Mutter, die um ihre Kinder fürchtet, schreit und ihnen nachläuft. [...] Bei dieser Interpretation könnten wir uns sogar vorstellen, dass ein Soldat den Kindern mit einer Geste signalisiert, sich fernzuhalten. (Ebd.)

Aufgabe 3

Erkläre, welche Fakten und Zusammenhänge die menschlichen Bodentruppen erkennen müssen, um die Personen im Haus als Zivilpersonen zu erkennen.

Aufgabe 4

Begründe, warum Emotionen für das Erkennen dieser Fakten und Zusammenhänge bei der *menschlichen* Interpretation der Situation entscheidend sind.

Exkurs: Isotropie

Isotropie bezeichnet die potenzielle Relevanz von allem für alles.

Was könnte der Preis für Tee in China mit Habibs Herzinfarkt zu tun haben?

Nun, das kommt darauf an. Wenn Habib stark in Unternehmen investiert hat, die Tee aus China exportieren, und Habib an einer Herzerkrankung leidet, die ihn anfällig für Herzinfarkte macht, wenn er unter enormem emotionalen Stress steht, und er erfährt, dass der Preis für Tee in China deutlich gefallen ist, was zu erheblichen Verlusten in seinem Portfolio führt und bei ihm ein hohes Maß an Stress und Angst auslöst, dann könnte es durchaus sein, dass der Preis für Tee in China relevant ist, um Habibs Herzinfarkt zu erklären.

Es ist schwierig, im Voraus, ohne die Einzelheiten einer Situation zu kennen, zu sagen, welche Informationen für die Begründung einer Behauptung oder einer Handlung relevant sein könnten und welche nicht.

Guarini & Bello 2011: S. 132

Aufgabe 5

Erläutere, warum Isotropie für das AWS im Sikh-Haus ein Problem ist und nicht immer *im Vorhinein* durch zusätzliche Informationen vermieden werden kann.

Aufgabe 6

[illegible]

Aufgabe 7

Entwirf einen verbesserten Vorschlag, wie man AWS so implementieren oder einsetzen könnte, dass auch Isotropie weitestgehend vermieden wird.

- ---

- ---

- ---

- ---

- ---

Objektivität – Ist ein AWS objektiver als menschliche Soldat(inn)en?

Jetzt hast du genug Hintergrundwissen über die Möglichkeiten und Probleme von KNN's und AWS erlangt, um die obige Leitfrage dieses Arbeitshefts angemessen beantworten zu können!

Aufgabe 1

Die behandelten Abschnitte sind alles Kriterien für ein objektives AWS. Ordne die folgenden Begriffe in die untere Hierarchie ein: ‚repräsentative Trainingsdaten‘, ‚Effektivität‘, ‚Transparenz‘, ‚Neutralität‘, ‚Situationsverständnis‘.

Zur Erinnerung: Dudeneintrag zu ‚objektiv‘:

1. tatsächlich (Bsp. objektive Tatsachen);
2. nicht von Gefühlen, Vorurteilen bestimmt; sachlich, unvoreingenommen, unparteiisch

- _____
 - _____
 - _____
 - _____
- _____
 - _____
 - _____

Um zu beantworten, ob AWS objektiver als menschliche Soldat(inn)en sind, musst du also zwei Teilfragen beantworten:

1. Sind AWS effektiver als menschliche Soldat(inn)en?
2. Sind AWS neutraler als menschliche Soldat(inn)en?

Aufgabe 2

a) Diskutiere stichpunktartig die Stärken und Schwächen von AWS und menschlichen Soldat(inn)en bei der Effektivität. Orientiere dich an die Unterpunkten von Effektivität aus der Hierarchie in Aufgabe 1.

AWS effektiver	menschliche Soldat(inn)en effektiver
+	—
+	—
—	+
—	+

b) Nimm Stellung zur Frage, ob AWS effektiver als menschliche Soldat(inn)en sind.

Aufgabe 3

a) Diskutiere stichpunktartig die Stärken und Schwächen von AWS und menschlichen Soldat(inn)en bei der Neutralität. Beziehe auch Transparenz als Unterpunkt mit ein.

AWS neutraler	menschliche Soldat(inn)en neutraler
+	–
+	–
–	+
–	+

b) Nimm Stellung zur Frage, ob AWS neutraler als menschliche Soldat(inn)en sind.

Aufgabe 4

Beurteile nun, ob AWS objektiver sind als menschliche Soldat(inn)en. Gehe dabei von einem realistisch möglichen AWS aus. Du darfst auch für eine Kombination aus AWS und Operator im Einsatz argumentieren. Wichtig ist, dass du deine Position begründest.

[illegible]

Verantwortung – Wer trägt die moralische Verantwortung für Kriegsverbrechen von AWS?

Selbst wenn du der Meinung bist, dass AWS objektiver als menschliche Soldat(inn)en sind, heißt das nicht automatisch, dass man AWS einsetzen sollte. Es gibt nämlich einige gewichtige Gegenargumente. Eines davon betrifft die moralische Verantwortung von Tötungen durch AWS. Dazu kommen wir jetzt.

Stellen wir uns vor, eine von einer hochentwickelten künstlichen Intelligenz gesteuerte Drohne bombardiert gezielt Soldat(inn)en, die eindeutig zu erkennen gegeben haben, dass sie sich ergeben wollen. Diese feindlichen Soldat(inn)en haben ihre Waffen niedergelegt und stellen keine unmittelbare Bedrohung für



Abbildung 6: Soldaten ergeben sich einer Drohne – Screenshot eines echten Drohnen-Einsatzes im Ukraine-Krieg

befreundete Streitkräfte oder Zivilpersonen dar. Es sei auch darauf hingewiesen, dass es sich bei diesem Bombenangriff nicht um einen Irrtum handelte; es gab keinen Zielfehler, keine Verwirrung bei den Befehlen der Maschine usw. Es war eine Entscheidung, die vom AWS in voller Kenntnis der Situation und der wahrscheinlichen Folgen getroffen wurde. Wir sollten in die Beschreibung des Falles einbeziehen, dass das AWS Gründe für sein Handeln hatte; vielleicht tötete es sie, weil es den militärischen Aufwand des Gefangennehmens als zu hoch ansah, vielleicht um die Gegner*innen in Angst und Schrecken zu versetzen, vielleicht um seine Waffensysteme zu testen [...]. Was auch immer die Gründe waren, sie waren nicht von der Art, dass sie die Tat moralisch rechtfertigten. Hätte ein Mensch die Tat begangen, wäre er sofort wegen eines Kriegsverbrechens angeklagt worden.

Wen sollten wir in einem solchen Fall wegen eines Kriegsverbrechens anklagen? Das [AWS] selbst? Die Person(en), die sie entwickelt haben? Den Operator, der ihren Einsatz befohlen hat? Niemanden? [...]

Sparrows Argumentation

Hinweis: Wenn hier die Prämissen (P) stimmen, muss auch die Konklusion (K) stimmen.

P4: Es kann niemandem – nicht den Entwickler(inne)n, nicht dem/der befehlshabenden Offizier(in) und nicht dem AWS – die moralische Verantwortung für ein Kriegsverbrechen, das von einem AWS begangen wurde, zugeschrieben werden.

P5: Wenn niemandem die moralische Verantwortung für Kriegsverbrechen, die von einem AWS begangen wurden, zugeschrieben werden kann, dann ist der Einsatz von AWS ethisch nicht vertretbar.

K: Also ist der Einsatz von AWS ethisch nicht vertretbar.

Aufgabe 1

Begründe, warum du dem Argument zustimmst oder ihm widersprichst, indem du die Prämissen prüfst.

Gegen die Verantwortung bei den Entwickler(inne)n

P1a: Ein AWS lernt über seine ursprüngliche Implementation hinaus aus seiner Umgebung und den eigenen Erfahrungen.

P1b: Wenn ein AWS über seine ursprüngliche Implementation hinaus aus seiner Umgebung und den eigenen Erfahrungen lernt, ist ein AWS nicht vollständig vorhersehbar.

P1c: Die Entwickler(innen) geben die Zuverlässigkeit des AWS dem Militär beim Verkauf mit an. (Methodische Annahme: Z. B. könnten sie angeben, dass das KNN im AWS Zivilpersonen zu 99,3% erkennt.)

P1d: Wenn ein AWS nicht vollständig vorhersehbar ist oder die Entwickler(innen) die Zuverlässigkeit des AWS dem Militär beim Verkauf mit angeben, kann den Entwickler(inne)n nicht die moralische Verantwortung für Kriegsverbrechen, die von ihrem AWS begangen wurden, zugeschrieben werden.

K1: Also kann den Entwickler(inne)n nicht die moralische Verantwortung für Kriegsverbrechen, die von ihrem AWS begangen wurden, zugeschrieben werden.

Aufgabe 2

(a) Begründe oder widerlege jeweils die drei Prämissen mithilfe deiner erlangten Kenntnisse.

P1a: _____

P1b: _____

P1d: _____

(b) Begründe stichpunktartig, ob du dieses Argument damit widerlegen kannst oder ob Sparrow sich gegen deine Einwände überzeugend verteidigen könnte.

Gegen die Verantwortung beim befehlshabenden Operator

P2a: Die Reaktionszeiten zum Eingreifen sind inzwischen zu kurz für den Menschen. (Überschallraketen erreichen 5km in 7s, Hyperschallraketen erreichen 5km in unter 1s.)

P2b: Wenn die Reaktionszeiten zum Eingreifen inzwischen zu kurz für den Menschen sind, muss die Entscheidung zur Tötung vom AWS ohne den befehlshabenden Operator getroffen werden.

P2c: Wenn die Entscheidung zur Tötung vom AWS ohne den befehlshabenden Operator getroffen werden muss, kann dem befehlshabenden Operator nicht die moralische Verantwortung für Kriegsverbrechen, die von einem AWS unter seinem Kommando begangen wurden, zugeschrieben werden.

K2: Also kann dem befehlshabenden Operator nicht die moralische Verantwortung für Kriegsverbrechen, die von einem AWS unter ihrem Kommando begangen wurden, zugeschrieben werden.

Aufgabe 3

(a) Begründe oder widerlege jeweils die drei Prämissen mithilfe deiner erlangten Kenntnisse.

P2a: _____

P2b: _____

P2c: _____

(b) Begründe stichpunktartig, ob du dieses Argument damit widerlegen kannst oder ob Sparrow sich gegen deine Einwände überzeugend verteidigen könnte.

Gegen die Verantwortung beim AWS

P3a: Die AWS empfindet kein Leid.

P3b: Wenn das AWS kein Leid empfindet, kann es nicht bestraft werden. (Wenn man versuchen würde, das AWS als Bestrafung zu beschädigen, ihm weniger Einsätze zu geben oder es zerstören würde, würde es kein Leid empfinden. Es wäre seltsam, das Strafe zu nennen.)

P3c: Wenn das AWS nicht bestraft werden kann, kann ihm nicht die moralische Verantwortung für Kriegsverbrechen, die vom ihm selbst begangen wurden, zugeschrieben werden.

K3: Also kann dem AWS nicht die moralische Verantwortung für Kriegsverbrechen, die von ihm selbst begangen wurden, zugeschrieben werden.

Aufgabe 4

(a) Begründe oder widerlege jeweils die drei Prämissen mithilfe deiner erlangten Kenntnisse.

P3a: _____

P3b: _____

P3c: _____

(b) Begründe stichpunktartig, ob du dieses Argument damit widerlegen kannst oder ob Sparrow sich gegen deine Einwände überzeugend verteidigen könnte.

Aufgabe 5

Beurteile, wer die Verantwortung für das Kriegsverbrechen, Soldaten, die sich ergeben haben, zu töten, tragen könnte. Beziehe die für dein Urteil entscheidenden Prämissen aus Aufgabe 2 – 4 mit ein.

[illegible]

Ausblick zu weiterführenden ethischen Debatten

*Wettrüsten um AWS verhindern – **Gefahr für die globale Sicherheit: Entwaffnungsschläge gegen taktische Nuklearwaffen** (vgl. P. Altmann 2019) **und Flugzeugträger möglich** (vgl. Hans Krech Drohnenexperte und Hauptmann d. R. 2023 im ZDF-Interview)*

„Und die Fähigkeit, von Anfang an, das mit reinzudesignen [...] Man wird dann sicherlich über die Zeit lernen, wie man die Schwellen sozusagen optimiert. Dass man am Ende nicht nur der Gute ist, der dann aber jedes Gefecht verliert, das automatisiert ist. [...] Aber wie man eben seine völkerrechtlichen und auch seine ethischen Standards dabei aufrechterhält.“ (Michael Schöllhorn vom Rüstungsunternehmen Airbus Defence and Space 2023 im ZDF-Interview)

AWS – eine gefühllose Waffe zur systematischen Massenvernichtung? Nie wieder Ungehorsam gegen Befehle. **Woran erinnert das ...?** (vgl. Renic & Schwarz 2023)

Nie wieder Kriegstote! AWS auf beiden Seiten für die ideale Kriegsführung!!! (vgl. Madock 2024)

Unsere Soldat(inn)en schützen uns, aber wer schützt Leib und Seele unserer Soldat(inn)en? (vgl. Burri 2017)

*„Und man muss bedenken, dass eine künstliche Intelligenz, die ja dann diese **Drohnenschwärme** steuert, **Menschen nicht als Menschen, sondern als Datenmuster betrachtet**. Und solche Systeme kennen z. B. auch **keinen Skrupel**. [...] Oder wenn sich ein **Fehler bei einer Drohne einschleicht z. B. wird der übertagen auf den ganzen Schwarm?**“ (Thomas Küchenmeister von der Campagne Stop Killer Robots 2023 im ZDF-Interview)*