

Hilfekarten

Trainingsdaten

2d:

1. Hilfekarte: Verändere die Distanz zwischen Objekt und Kamera.
2. Hilfekarte: Probiere verschieden aussehende Objekte für die gleiche Klasse aus.
3. Hilfekarte: Teste ein Objekt, das man keiner deiner Klassen zuordnen kann.

3d:

1. Hilfekarte: Man kann während des Trainings nach jeder Epoche das KI-Modell auf den Testdaten überprüfen und die acc und test acc überprüfen und sieht dann den Verlauf des rechten Diagramms bis zur aktuellen Epoche.
2. Hilfekarte: Kann man irgendwie erkennen, wann das KI-Modell anfängt zu overfitten?
3. Wird das KI-Modell für immer desto besser, je länger man es auf den Trainingsdaten trainiert?

Effektivität

5b

1. Hilfekarte: Überlege dir, welches Problem entstehen würde, wenn man ein KNN auch mithilfe der Testbilder trainieren würde.
2. Hilfekarte: Probiere aus, was passiert wenn man ein KNN mit Testbildern trainiert. Kopiere alle Testbilder in den Ordner für die Trainingsbilder, trainiere dann ein neues Modell damit und teste es.
3. Hilfekarte: Warum ist dieses bessere Ergebnis ein Problem, wenn man das Modell jetzt in der realen Welt auf nochmal neue Bilder einsetzt?

Lösungsskizzen

Lösungsskizze Einstieg

1.

- Helden, die Iron Man und seinen Freund retten und „die Bösen“ eliminieren; effektive und schnelle Unterstützung für menschliche Soldaten

2.

- erst präzise sehr effiziente Waffen, die keine unnötigen Gefühle haben, die sie vom Kämpfen abhalten;
- am Ende gruselig/abschreckend: töten Aktivisten/Studenten wegen Hashtag; andersdenkende Politiker (schlecht für das Allgemeinwohl in falschen Händen)

3.

1. Video:

- Helden, Unterstützer (gut für das Allgemeinwohl)
- Zuschauer fiebern mit Tony Stark mit und freuen sich über Auftritt von AWS

2. Video

- Zuschauer sollen zu Beginn fiktivem Moderator folgen, dann abgeschreckt werden von den fiktiven Nachrichtenszenen und dann vom Beifall des fiktiven Publikums verwirrt werden und dann kritischem Statement von KI-Professor, der gegen komplett autonome AWS ist, folgen

➔ verschiedene Formate: heroische Filmszene vs. Kurzfilm mit Statement, der abschrecken soll

4. Mindestens 2 von diesen oder 2 gleichwertige andere Argumente

- Autonome Waffensysteme haben keine Gefühle, die ihr Urteilsvermögen trüben
→ Präzise/effektiv, aber Ziele von Menschen, die sie einsetzt, abhängig (Allgemeinwohl vs. Missbrauch für ideologische Zwecke oder Diktatur)
- Maschinen sollten nicht selbst über Leben und Tod entscheiden: Hashtags sagen nicht genug über Menschen aus, um effektiv Terroristen über Ideologie sicher zu identifizieren, Unschuldige könnten getötet werden
- dass Menschen sich weigern, kann auch etwas sehr Positives sein, erhabene Ziele sind ethische Prinzipien und sollten keinesfalls ignoriert werden
→ auch in Kriegen schießen laut Studien viele Soldaten absichtlich daneben oder verbrachten viel Zeit mit nicht notwendigem neu laden der Waffen, sodass es insgesamt weniger Tote und Verletzte gab; AWS würden effizienter und mehr töten, zerstören und verletzen

Lösungsskizze Repräsentativität (Was sind gute Trainingsdaten?)

2.

c)

- Batch-Size: Trainingssamples (Teilmenge der Trainingsbilder) pro Lernschritt
- Epochen: Anzahl der Wiederholungen aller Trainingsbilder beim Lernen

3.

b)

- Distanz von Objekten variieren
- Perspektive/Winkel ändern
- Schatten auf Objekt
- anderes Licht
- Teil der Objekte verdecken (auch bei unwesentlichen verdeckten Details)
- Objekte einer der eigenen Klassen wählen, das sich stark von Trainingsbildern unterscheidet (z. B. bei Klasse Stuhl: Drehstuhl)
- Objekt, das keiner der Klassen angehört erkennen lassen → erkennt fälschlicherweise eine der Klassen mit hoher Sicherheit
- mehrere Objekte auf dem Bild verwirren KI-Modell

c)

1. möglichst viele Bilder (damit unwesentliche Details keine zu hohe Rolle spielen: z. B. Sperrbildschirm, um Handy zu erkennen)
2. möglichst verschiedene Bilder (Diversität)
 - a. Perspektive variieren
 - b. Distanz variieren
 - c. alle stark unterschiedlichen Unterarten
 - d. Licht/Schatten integrieren
3. eigene Fehlerklasse und Trainingsbilder für Objekte, die keiner Klasse entsprechen

d)

- falsche Vorhersagen aufgrund von falsch gelernten Mustern
 - insbesondere: Muster unvollständig in Trainingsdaten
 - irrelevante Gemeinsamkeiten in Daten werden überproportional wichtig

4.

a) 100% militärische Daten:

das AWS kann beispielsweise verschiedene Panzertypen voneinander unterscheiden, aber keine Autos erkennen → könnte Auto für Panzer halten oder auch Polizist für Soldat halten (out of distribution Problem) und abgeschossen werden, Tötung von Zivilpersonen oder Polizisten

b) 90% militärische Daten: 10% zivile Daten ohne Antennen auf ziviler Gebäuden: Antenne auf zivilem Gebäude könnte als militärische Radaranlage interpretiert und beschossen werden, wobei viele Zivilpersonen sterben könnten

c) 50%:50%, wobei militärische Daten von neuen Technologien hauptsächlich verpixelte Satellitenaufnahmen aus großer Distanz sind: ´

besser als vorherige Szenarien, aber nicht ideal; gerade bei verpixelten Bildern unklar, was gelernte Muster sein werden → verpixelt bedeutet oft Verlust von Details, also schlechteres Lernen von Mustern → Gefahr von Nichterkennung oder Fehlerkennung; wenn aber zivile Gebäude, Fahrzeuge und Personen zuverlässig erkannt werden weniger problematisch

(Bsp. Erkennung 99,3% als Zivilflugzeug und 99,98% als neuer russischer Kampfjet sollte man trotzdem nicht abschießen)

5.

a)

- Accuracy links: je mehr Epochen, desto besser lernt KI-Modell Muster

b) mögliche Optionen

- wird zwischenzeitlich schlechter, weil es verschiedene Muster ausprobiert, dann richtige Muster und wird nur aus Zufall extrem gut auf Trainingsdaten, was dann korrigiert wird
- lernt irgendwann mehr unwichtige Muster der Trainingsdaten, wenn Daten nicht mehr hergeben

c)

- Overfitting: Auswendiglernen der Trainingsdaten (zufällige eigentlich unwichtige Muster werden mitgelernt, je länger auf den eingeschränkten Trainingsdaten gelernt wird)

d) Lösung für Overfitting aus Diagramm

- Kurven von Vorhersagen auf Trainingsdaten und Testdaten vergleichen: wenn nicht mehr besser auf Testdaten wird, dann abbrechen

e)

- AWS würden sich zu stark an Trainingsdaten orientieren und wären überfordert in komplexen unvorhergesehenen Situationen auf dem Schlachtfeld
- Beispiel: würden Zivilpersonen nur erkennen, wenn sie Haare haben oder Kinder sind oder alle Frauen sind Zivilpersonen statt auf Waffen zu achten

Lösungsskizze Effektivität

1.

- (a) Input: Farb- oder Helligkeitswerte für jedes Pixel des Bildes, das klassifiziert werden soll
- (b) Output: Klasse (in diesem Fall 0, 1, 2, 3, 4, 5, 6, 7, 8 oder 9)
- (c) Hidden Layer: Verarbeitung des Bildes, Erkennen von Mustern
- (d) Bias: Beeinflussung des Output-Wertes
- (e) Gewichte: Relevanz von bestimmten Neuronen verstärken/abschwächen
- (f) Aktivierungsfunktion: bestimmt Schwelle, ab der ein Neuron seinen Wert weitergibt

2.

- (c) Trainingsstatistik: 72%, 65%, 69%, 71%, 70%
Teststatistik: 75%, 69%, 74%, 74%, 73%

3.

- (a) 1. Input-Neuron links oben,
2. Input-Neuron rechts oben,
3. Input-Neuron links unten,
4. Input-Neuron rechts unten

Erklärung: Schwarz-Weiß-Werte hängen mit Neuronen der Eingabeschicht zusammen, Bilder bestehen aus 4 Pixeln:

Je heller oder transparenter ein Neuron, desto dunkler ist die Farbe (bei Farbinformation 0) und je dunkler ein Neuron (Farbinformation näher an 1), desto heller die Farbe.

- (b) 1. Output-Neuron: dunkel
2. Output-Neuron: grau
3. Output-Neuron: hell

Je blauer das Output-Neuron, desto höher ist sein Wert und umgekehrt. Wenn also das Bild hell ist, wird das Hell-Klasse-Neuron am blauesten sein. Das gleiche gilt für die beiden anderen Output-Neuronen für dunkel und grau.

4.

a)

KI-generiert (bearbeitet): ChatGPT basierend auf <https://databasecamp.de/ki/backpropagation-grundlagen>

Das neuronale Netz betrachtet eine handgeschriebene **2**. Das Bild besteht aus 784 Bildpunkten, die als Zahlen an die erste Schicht des Netzes weitergegeben werden. Von dort wandern die Informationen durch viele Neuronen und Schichten. Jedes Neuron verarbeitet die Werte mit seinen Gewichten, Biases und der Aktivierungsfunktion. Am Ende trifft die Ausgabeschicht ihre Vorhersage: „Das ist wahrscheinlich eine 3 oder eine 6.“ Dabei war die richtige Antwort eigentlich die **2**. Das Netz hat sich also deutlich geirrt.

c) 0,8

d) Biases, Gewichte

e)

KI-generiert (bearbeitet): ChatGPT basierend auf <https://databasecamp.de/ki/backpropagation-grundlagen>

Aber wo genau ist der Fehler entstanden?

Jetzt wird es schwieriger. Das Netz besitzt viele Schichten mit unzähligen Verbindungen. Wenn am Ende die falsche Zahl herauskommt, ist zwar klar, dass das Netz einen Fehler gemacht hat und auch wie groß der Fehler ist (bei unserer Vorhersage von einer 2 mit einer Vorhersage von 0,2 wäre das 0,8 daneben). Es ist aber nicht sofort klar:

Wo genau ist der Fehler? Welches Neuron hat den Fehler verursacht? Vielleicht hat schon eine frühe Schicht die Form der „2“ falsch erkannt. Vielleicht entstand der Fehler aber erst ganz am Ende. Und noch etwas macht die Sache kompliziert: Selbst wenn sich ein Neuron „verrechnet“, muss das gar nicht wichtig sein. Denn die nächste Schicht könnte diesen Wert fast ignorieren — zum Beispiel, weil die Verbindung ein Gewicht nahe 0 besitzt. Das bedeutet: Wie stark ein Neuron zum Fehler beiträgt, hängt auch davon ab, wie die nachfolgenden Schichten mit seinem Ergebnis umgehen.

f)

KI-generiert (bearbeitet): ChatGPT basierend auf <https://databasecamp.de/ki/backpropagation-grundlagen>

Kurz zusammengefasst

Das Netz

1. macht eine Vorhersage,
2. überprüft das Label und misst den Fehler,
3. verfolgt den Fehler rückwärts durchs Netz mithilfe von Backpropagation,
4. findet heraus, welche Gewichte und Biases wie stark verantwortlich waren,
5. und verändert sie in Richtung des steilsten Abstiegs (Gradientenabstieg).

3. und 4. können auch zusammengefasst werden

5.

- (c) Overfitting: Netz lernt irrelevante Details aus Trainingsdaten auswendig
- (d) Die Trennung zwischen Trainingsdaten und Testdaten ist wichtig, damit überhaupt erkennbar ist, wann das Netz overfitted ist, also nur noch auswendig lernt
Normalerweise wird das Netz irgendwann auf den Testdaten nicht mehr besser, während es immer noch besser auf den Trainingsdaten wird. Aber wenn man es auf den Testdaten trainiert, wird es auch immer besser auf den Testdaten, weil es diese auch noch auswendig lernt. Durch mehr Daten können zwar besser Muster erkannt werden, aber dann kann man nicht mehr erkennen, wann das Netz overfitted wird und damit auf der Realität eigentlich wieder schlechter beim Vorhersagen wird.

6.

Effektivität eines AWS beim Unterscheiden zwischen Wohnblock und Militärbasis:

a) Anzahl von Neuronen und Hidden Layers:

genügend Neuronen und Layer, um komplexe Muster in Form von Biases und Gewichten überhaupt theoretisch abbilden zu können: Ein einzelnes Haus reicht vielleicht nicht immer zur Erkennung, ob es ein Wohnblock oder eine Militärbasis ist, vielleicht erst das Zusammenspiel aus mehreren Gebäuden und Anlagen

b) Biases und Gewichte (durch Training)

geeignete Biases und Gewichte für konkrete Muster → Je geeigneter, desto effektiver das KNN → Stacheldrahtzaun, Schießplatz etc. muss als solcher erkannt werden können

c) Epochenanzahl

Je mehr Epochen, desto mehr Details werden gelernt → bessere Muster z. B. Stacheldrahtzaun, Wachtürme, herumfahrende Panzer, Schießplätze, aber auch Overfitting durch Lernen von irrelevanten Mustern in Trainingsdaten (bei repräsentativen Trainingsdaten weniger)

d) Aufteilung der Daten in separate Trainings- und Testdaten:

Repräsentativität der Daten kann im Nachhinein besser beurteilt werden bei Aufteilung: keine Aufteilung → mehr Daten und Muster zum Lernen, aber keine sinnvolle Überprüfung möglich: z. B. Stacheldraht erkennbar, wenn er nicht 1:1 wie auf dem trainierten Bild aussieht oder anders befestigt ist oder der Schießplatz relativ gesehen zu den Eingängen woanders ist?

Zusatzaufgabe

(b)

Lernrate zu klein: Netz verbessert sich nach einer Zeit nicht mehr, kommt nicht aus einem hoch gelegenen Tal heraus weiter nach unten

Lernrate zu groß: Netz wird instabil, verbessert und verschlechtert sich andauernd schlagartig, findet meist nicht tiefes Tal, sondern irgendein zufälliges Tal, was für die aktuelle mitunter sehr spezifische Batch geeigneter ist

Lösungsskizze Transparenz

Aufgabe 1

beide Klassifikationen sind mehr oder weniger richtig:

- a) Haus wird an Haus erkannt, an Plateau und Zaun am Ufer (hauptsächlich, wobei aus unerfindlichen Gründen die Ecken des Bildes auch relevant waren)
- b) Auto wird an den Autos erkannt (auch die Leerräume dazwischen hatten anscheinend eine gewisse Relevanz)

Aufgabe 2

- Wasserzeichen von Pferdeliebhhaber → meistens Pferde-Fotos
- Meer um Kreuzfahrtschiff mit Wellen → wahrscheinlich Schiff (Bilder von Meer mit Wellen seltener ohne Schiff in Datenbank)

Aufgabe 3

- a) Clever Hans war Pferd, von dem alle dachten, dass es rechnen könnte; in Wirklichkeit hat es nur die Mimik und Gestik des Fragestellers gelesen und wusste so, wann es aufhören musste mit den Hufen zu klopfen → hatte keine Ahnung vom Rechnen
- b) KNN können sich solche Tricks auch zunutze machen und so schnell vermeintliche Muster finden, die aber bei echten Daten nicht mehr funktionieren oder unzuverlässig sind (heuristische Methoden)

Aufgabe 4

a)

- kleine Brille spricht für Klassifikation ‚Mann‘ und gegen ‚Frau‘
- große Brille spricht für ‚Frau‘
- glänzende oder farbenfrohe Lippen sprechen für ‚Frau‘
- kurze Haare und Bart sprechen für „Mann“

b)

- Stereotype: ‚Frauen haben lange Haare und Männer kurze Haare‘ und ‚Frauen tragen Lippenstift und Männer nicht‘
- KI-Modell basiert auf Datensatz, den Menschen gelabelt haben und wiederum andere Menschen für repräsentativ genug gehalten haben

c)

- auch bei perfekten Labeln innerhalb von Trainingsdaten kann die Auswahl an Trainingsdaten und die Auswahl von zur Verfügung stehenden Labeln Stereotype repräsentieren
 - auch Frauen können kurze Haare oder eine kleine Brille haben (kam in Trainingsdaten für ‚Frau‘ wohl nicht oder zu wenig vor)
 - Männer können auch glänzende oder farbenfrohen Lippen haben (kam in Trainingsdaten für ‚Mann‘ wohl nicht oder zu wenig vor)
 - Transpersonen werden vermutlich nicht berücksichtigt und würden dann oft falsch erkannt werden
 - Nonbinarys über Label ‚Mann‘ und ‚Frau‘ hinaus wie Hermaphroditen fehlen

Aufgabe 5

- Overfitting: nach langer Zeit, wenn aus Trainingsdaten nichts mehr herauszuholen ist, und die richtigen Muster schon gelernt wurden, werden mehr und mehr unwichtige Details auswendig gelernt
- Clever-Hans-Problem: schnelle relativ hohe Sicherheit des KNN's, aber ignoriert richtige Muster vielleicht sogar komplett (wie bei Boot), vor allem wenn Daten so begrenzt, dass sie nichts Besseres hergeben, weil zufällige Gemeinsamkeiten in Trainingsdaten bestehen
- technisch mit Berglandschaft:
 - Clever Hans-Problem: erstbestes Tal (relativ niedriger Fehler),
 - Overfitting: sucht in der Nähe nach niedrigstem Tal, aber orientiert sich dabei zu sehr auf Trainingsdaten und findet dafür noch ein immer niedrigeres Tal durch Lernen immer irrelevanterer Muster, was nicht am niedrigsten für die Testdaten ist

Aufgabe 6

a)

- AWS lernt aus begrenzten Trainingsdaten, dass Frauen nur Rucksäcke mit Tarnfarben tragen, wenn sie Aufständische/Soldatinnen sind und zivile Frauen keine Rucksäcke mit Tarnfarben tragen können → Abkürzung mit schnell mit meist guter Vorhersage
- aber kein angemessener Grund ist, um sie als Aufständische zu klassifizieren (kein logischer oder empirisch notwendiger Zusammenhang)

b)

- AWS lernt Muster und hat insofern Gründe für Klassifizierungen; Operator im Unklaren über Gründe lassen fördert unnötige Verwirrung
- Klassifikationen entscheiden hier über Leben und Tod von Menschen: falsch als Soldat(in)/Aufständische(r) klassifizierte Zivilpersonen werden getötet, was ein Kriegsverbrechen darstellt
- AWS hat keine 100%ige Test Accuracy bei Zivilpersonen, Gefahr der Fehlerkennung besteht also → können nur Operator verhindern, und zwar am besten, wenn sie die Gründe für die Klassifikation nachvollziehen können

c)

- Video-Live-Stream enthält tausende von Einzelbildern, wobei die relevanten Muster vielleicht nur auf manchen Bildern davon zu erkennen sind → Zeit ist zu knapp, um alle zu analysieren
- falsche Muster sind nicht immer an einem einzelnen Bild zu erkennen, sondern manchmal erst, wenn man sich mehrere falsch klassifizierte Bilder ansieht
- Muster, die KNN erkennt, könnten sogar richtig sein, aber Menschen können manchmal nichts mit der Saliency Map anfangen (zu unscharf oder unbekanntes aber richtiges Muster möglich) → könnten dem KNN irgendwann blind vertrauen, wenn es sich herausstellt, dass es in diesen Fällen meistens recht hat
- Bestätigungsfehler: Suche nach zusätzlichen Indizien, die AWS-Klassifikation bestätigen, weil es den vermeintlichen Waffenlauf nur nicht richtig erkannt hat

Lösungsskizze Neutralität

1. z. B. 5 aus folgender Liste:

- Einsatz
 - Mensch überwacht alle Tötungs- und Zerstörungsentscheidungen
 - AWS schlägt vor und Mensch muss bestätigen ODER zumindest Mensch hat Möglichkeit zu widersprechen
 - zivile Umgebungen meiden (Einsatz nur auf hoher See oder unbewohnten Orten)
 - begrenztes Einsatzgebiet definieren
 - konkrete Ziele vorher aufklären und durch Menschen bestätigen, sodass Ziele im Tötungseinsatz nur bestätigt werden müssen
- Implementierung/Design
 - Die Gründe für jede Tötungs- und Zerstörungsentscheidung müssen transparent nachvollziehbar sein. Z. B. Saliency Map mit aussagekräftigsten True Positive und False Positive Bildern
 - Ethics Module, worin alle gängigen Stereotype eingespeist sind, überprüft, ob in aktueller Situation Entscheidung durch Stereotype kompromittiert sein könnte → wird nicht ausgeführt, außer Mensch überschreibt den Befehl des AWS aktiv
 - viel Aufwand für gute Aufklärung: näher heranfliegen für Nahaufnahmen, technische Möglichkeiten ausschöpfen, insb. Informationen aus Datenbanken nutzen, Infrarot usw.
 - nur bei extrem hohem Vorhersagewert oder expliziter Auswahl von Menschen Gewalt ausüben
 - viele mögliche schwierige Kampfsituationen von Experten einholen und in Design berücksichtigen

2.

- sehr viele behandelten Zivilpersonen nicht mit Respekt: wendeten sinnlos Gewalt an oder zerstörten sinnlos deren Eigentum und würden sie zur Erlangung von Informationen foltern, manche taten es auch ohne irgendeinen Nutzen
- sehr viele deckten sich gegenseitig statt ethische Verstöße zu melden
- Erfahrung auf dem Schlachtfeld und der Verlust von Teammitgliedern führte gehäuft zu Wutanfällen und noch mehr ethischem Fehlverhalten: emotionale Instabilität

3.

a)

- Emotionalität hemmt beim Töten
- erfahrene und unerfahrene Truppen haben die gleiche Hemmung beim Töten
- Soldat(inn)en versuchen Gewalt sogar trotz Befehlen zu vermeiden, indem sie ihre Zeit mit sinnlosem Laden verbringen statt zu schießen

b)

- Emotionalität (Wut, psychische Probleme) führt zu doppelt so vielen Tötungen vs. Emotionalität hemmt beim Töten (uneingestandene Hemmung)
- erfahrene Truppen haben weniger Hemmung beim Töten (erlebte Verluste → Rache) vs. erfahrene und unerfahrene Truppen haben die gleiche Hemmung vor dem Töten
- viele Soldat(innen) Verletzen unnötig Zivilpersonen und beschädigen deren Eigentum und noch mehr decken Kriegsverbrechen anderer vs. Soldat(inn)en versuchen alles, um Gewaltausübung zu vermeiden

c) Wie beide Darstellungen gleichzeitig wahr sein können: beispielhafte Optionen

- US-Militär nahm die Forschungen des Historikers so ernst, dass sie Gegenmaßnahmen ergriffen und die Soldaten bei der Ausbildung so brechen und mit sinnloser Gewalt konfrontieren, dass sie deutlich weniger Hemmungen beim Töten haben und auch die Achtung vor Zivilpersonen und deren Eigentum verlieren
- hier werden Äpfel mit Birnen verglichen
 - frühere Kriege fanden nicht zwischen Berufssoldaten statt, sondern meist zwischen wehrpflichtigen Bürgern, die eine höhere Gewaltabneigung haben
 - ein normaler Krieg und eine Aufstandsbekämpfung sind vollkommen verschieden, weil Soldaten bei letzterer darauf konditioniert werden jeder Zivilperson zu misstrauen und sie für einen potentiell gefährlichen Aufständischen zu halten: sie sehen ja tatsächlich genauso aus, wenn sie keine Waffe tragen, denn keine Uniform, keine Abzeichen
- 1. Bericht vom US-Militär zwar wahr, aber verzerrte Interpretation: Soldaten decken sich gegenseitig → Sympathie, Kameradschaft, Pflichtgefühl gegenüber eigenem Team, was einem vielleicht schon oft das Leben gerettet hat, Mitleid; Soldaten wissen in einigen Situationen nicht, wie sie sich ethisch verhalten sollen → Könnte auch etwas Gutes sein, wenn sie die Komplexität eines moralischen Dilemmas überhaupt erkennen können.

- die Aufstandsbekämpfung hat so viel Wut auf vermeintliche Zivilpersonen erzeugt durch gefallene Teammitglieder, dass Hemmung zum Töten keine Rolle mehr spielt (nicht repräsentativ)

4.

- Menschen
 - können Vorurteile mit Stereotypen haben
 - emotional: können bei Verlust von Teammitgliedern oder Freunden emotional instabil werden → vermehrt ethische Verstöße möglich
 - im Normalzustand meist eher Abneigung gegen Gewaltausübung
 - kameradschaftlich im Guten wie im Schlechten
 - wissen wenigstens, wenn sie Amok laufen
 - zweifelt daran, wie man sich ethisch richtig verhalten kann
 - haben früher oder später eigene Kampferfahrungen, wenn sie im Kampf eingesetzt werden
- AWS
 - emotionslose Maschine: kann durch Stereotype aus Trainingsdaten oder der Auswahl an Daten kompromittiert sein, kann aber durch Überwachung ausgeglichen werden
 - keine Hemmungen vor Gewalt: weiß nicht, wenn es Amok läuft → überwacht durch Menschen, der in Sicherheit ist und nicht Gefahr läuft Kameraden zu verlieren, sondern nur ersetzbares AWS
 - zweifelt nicht daran, wie man sich ethisch richtig verhalten kann
 - kein Motiv für Gräueltaten
 - keine eigene Kampferfahrung, nur fertig klassifizierte Daten aus fremden Kampfszenarien
 - vorsichtig: nur bei sehr hohem Vorhersagegrad Gewaltanwendung; nur in bestimmten Umgebungen überhaupt zulässig (z. B. nicht bei Aufstandsbekämpfungen in Städten)
- Gemeinsamkeiten
 - Stereotype können Entscheidungen beeinflussen
 - können unbewusst Fehler machen
 - können Amok laufen

5. Ja,

- wenn AWS eingeschränkter sind, was Umgebung angeht
- vorsichtiger
- werden nicht emotional instabil, aber haben keine Hemmungen gegen Gewalt
- können transparent von Menschen überwacht und kontrolliert werden → Mensch dann nicht in Gefahr und verliert auch keine Teammitglieder, wodurch emotionale

Instabilität oft vermieden werden könnte und Mensch erkennt, wenn AWS Amok läuft, kann es stoppen

- Menschen sind nützlicher als Operator. Können immer noch ethisch richtige Entscheidung hinterfragen und Abneigung gegen Gewalt und Kampferfahrung einbringen.

Lösungsskizze Situationsverständnis

1.

- eliminiert alle potentielle Ziele, die mit hoher Sicherheit als Aufständische vorhergesagt wurden, wenn sie sich nicht rechtzeitig ergeben

2.

- Waffen besonders rot, ebenso ggf. bestimmte Erkennungskleidung/Abzeichen

3.

- potentielle Ziele sind Kinder → sehr jung
- sie spielen gerade → Ball
- anderes potentielles Ziel Mutter → Art der Interaktion (besorgt um Kinder)
- Mimik der Kinder und der Mutter → Emotionen → Intentionen nicht feindlich
- Kirpan als religiöses Symbol → Dolch keine Waffe → Kirpan-Träger kein Ziel

=> enorme Menge an Hintergrundwissen und Kontextwissen über Kinder, Familien, Bälle und Sikh's erforderlich

4.

- Mutter neben ihren Kindern will sie bloß schützen und würde für andere Zwecke normalerweise niemanden vor den Augen ihrer Kinder angreifen
- Kinder sind fast immer harmlos gegenüber Soldaten oder AWS: Kinder zu töten könnten sich die meisten menschlichen Soldaten niemals verzeihen (selbst bei Kindersoldaten hätten sie wohl Probleme)

→ intuitive Zurückhaltung

5.

- der zentrale Punkt ist: Woher weiß ich, welche Informationen relevant sind und welche nicht?
- Ist ein Dolch relevant dafür ein Aufständischer zu sein? normalerweise schon, aber nicht notwendigerweise; ein Kirpan spricht dagegen, außer ein Sikh läuft Amok und benutzt sein Kirpan als Waffe; auch ein Jagdmesser muss nicht dafür sprechen, kann es aber, gerade bei Aufständischen im Vergleich zu professionellen Soldaten möglich
- Kindersoldaten gibt es, aber meist spricht der Fakt, dass es ein Kind ist, dagegen, dass es ein Soldat ist.
- Welche Informationen relevant sind, um jemanden z. B. als Aufständischen zu erkennen, muss ein Mensch oft anhand der Einzelheiten in der Situation neu

beurteilen; ist nicht allgemein gültig und immer gleich; AWS können das (noch) nicht

6.

a)

- Isotropie betrifft beide – sowohl AWS als auch menschliche Soldaten: auch Menschen können Kirpans fälschlicherweise als gefährliche Waffen an Kindersoldaten deuten
- aber bei AWS Isotropie noch problematischer wegen Abwesenheit von Emotionen und damit verbundener emotionaler Intelligenz mit sozialem Kontextwissen → intuitive Zurückhaltung; Stärke von Menschen (übersteigt Fähigkeiten von reiner Bildklassifikation)
- Menschen sind gestresst und können deswegen Fehler machen, AWS nicht

b)

- Nein, Menschen können zwar gestresst sein, aber in der Regel sind sie trotzdem so geübt darin und über Emotionen dazu konditioniert, Situationen zu erfassen und mit sozialem Kontextwissen zu verstehen, dass sie Situationen insgesamt besser verstehend als AWS.

Aufgabe 7 (mögliche Ansätze)

- Einsatz
 - von Menschen überwachen lassen (Autonomie beschränken): Mensch muss bestätigen oder kann eingreifen (über Saliency Maps hinaus hohe Transparenz ermöglichen)
- Implementierung
 - sicherstellen, dass Menschen auch bei Fehlinformationen aus Datenbanken das Recht zum Überschreiben von Tötungs-Entscheidungen haben
 - Social Encoder Modul integrieren: interpretiert Mimik, Gestik und interpersonale Zusammenhänge; bezieht psychologisches, kulturelles, religiöses und ethnisches Kontextwissen mit ein
 - KI mit Allgemeinwissen trainieren (Mischung aus etwas wie ChatGPT und AWS) implementieren
 - Trainingsdaten überprüfen/filtern auf Stereotype
 - Gedächtnis, um aus Fehlern (nicht unbedingt Tötungen, kann auch nur Widerspruch zur Tötung von Operator sein) zu lernen, Widersprüche müssten aber im Nachhinein begründet werden
 - für geringe Vorhersage-Sicherheit bei Objekten Out of Distribution (keiner bekannten Klasse zuordenbar) sorgen

Lösungsskizze Objektivität

Aufgabe 1

- Objektivität erfordert:
 - Effektivität bei Unterscheidung zwischen feindlichen Soldat(inn)en und Zivilpersonen → objektive Tatsachenfeststellung
 - repräsentative Trainingsdaten
 - gutes Training
 - Situationsverständnis
 - Neutralität (Unvoreingenommenheit)
 - dafür Transparenz nötig

Aufgabe 2 und 3:

- Pro-Contra Liste als Zusammenfassung der Erkenntnisse aus vorherigen Abschnitten:
 - Effektivität
 - AWS effektiver:
 - + durch möglichst repräsentative Trainingsdaten und gutes Training können KNN's von AWS subtilere Muster erkennen, ggf. auch in verpixelten Bildern → mehr true positives
 - können aber auch falsche Muster sein bei Overfitting oder Mangel an bestimmten Daten
 - Mensch effektiver:
 - + Situationsverständnis dank Emotionen extrem zuverlässig im Vergleich zu AWS; müsste erst mal nachgeahmt werden; erst in ferner Zukunft könnte ein AWS den Menschen hier übertreffen → weniger false positives
 - durch Stress können sie aber beeinträchtigt werden
 - → Mensch und AWS in verschiedenen Bereichen jeweils besser, keiner klar überlegen (in naher Zukunft)
 - Neutralität
 - AWS neutraler:
 - + Fehlen von Emotionen oder bewussten Stereotypen → keine absichtlichen Gräueltaten
 - aber auch keine Hemmungen vor dem Töten
 - Auswahl an Trainingsdaten kann menschliche Stereotype für Klassifikation trotzdem relevant machen und diese kompromittieren
 - Mensch neutraler:
 - + hat emotionale Hemmung vor dem Töten, sofern nicht abtrainiert

- + kann bewussten Stereotype entgegenwirken, wenn er will; hat Fähigkeit zur Reflexion
- wegen unbewusster Stereotype und Wut bei Verlust von Teammitgliedern können sie Zivilpersonen unnötig Schaden
- → KNN's sind nicht automatisch neutral oder neutraler als menschliche Soldat(inn)en, muss wegen versteckten Stereotypen in Trainingsdaten und Clever Hans Problem immer wieder überprüft werden von Entwickler(inne)n und auch Operators; diese erleben keine Verluste von Teammitgliedern wegen sicherer Distanz vom Schlachtfeld → Notwendigkeit von Transparenz und menschlicher Überwachung bei Einsätzen (zumindest solange KNN's nicht neutralere Ergebnisse als Menschen liefern)

Aufgabe 4

- AWS könnten irgendwann objektiver werden als menschliche Soldat(inn)en, ist aber aufgrund von Problemen nicht absehbar
- Effektivität (als test accuracy):
 - höher, wenn Fehlerkennungen (false positives) vermieden werden können;
 - Clever Hans Probleme sind gravierender Einschnitt für Effektivität, weil Menschen das so nicht passieren würde (offensichtliche Fehler) → durch rationaleres logisches Denken vielleicht vermeidbar
 - Situationsverständnis schwierig ohne extrem gutes social encoding Modul: unklar, wann so etwas möglich ist
 - Menschen können durch Emotionalität vor falschen, aber auch legitimen Tötungen zurückschrecken, können aber auch durch Verlust Wut empfinden und Gräueltaten z. B. an Zivilbevölkerung begehen
- Neutralität:
 - ähnlich bei AWS und Menschen: bei AWS kompromittierte Trainingsdaten möglich und bei Menschen Emotionen und Stereotype, die Neutralität einschränken
 - kompromittierte Daten müssten irgendwie erkannt und herausgefiltert werden → sogar über Clever Hans Problem hinaus, wenn Daten objektive Kriterien für Erkennung gar nicht mehr hergeben
- AWS noch nicht objektiver: muss genannte Probleme überwinden

Lösungsskizze Verantwortung

1.

- P4: Verantwortung doch zuschreiben mit folgenden Optionen:
 - Entwickler: haben fehlerhafte autonome Tötungsmaschine hergestellt
 - Offizier: hat nicht aufgepasst und zugelassen, dass AWS Soldaten tötet, die sich bereits ergeben haben
 - AWS: hat autonom gehandelt und ist allein kausal für die Tötung verantwortlich, also auch ethisch; sollten niemals so frei eingesetzt werden
- P5 schwer zu widersprechen
- → Widerspruch gegen Argument oder alternativ Zustimmung, dass AWS verboten werden sollten

2.

(a)

- **P1a:** Ein AWS lernt neben seiner Programmierung auch aus seiner Umgebung und den eigenen Erfahrungen.
→ AWS kann fertig ausgeliefert werden, ohne dass es im Einsatz lernt und sein Verhalten verändert; es könnte für weiteres Lernen und Anpassung an die Einsatzbedingungen von den Entwicklern gewartet werden, sodass es nicht im Einsatz lernt
- **P1b:** Wenn ein AWS neben seiner Programmierung auch aus seiner Umgebung und den eigenen Erfahrungen lernt, ist es nicht vollständig vorhersehbar.
→ Selbst wenn es nicht aus seiner Erfahrung lernt, ist es nicht wirklich vorhersehbar, was im Einzelfall bei einem Input-Bild passiert. Ein unentdeckter Clever-Hans-Trick oder Objekt unbekannter Klasse, fehlendes Situationsverständnis durch Isotropie oder leichtes Overfitting könnten Ursachen dafür sein
- **P1d:** Wenn es 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.
→ Man könnte widersprechen und behaupten, wenn es Risiken gibt, dass Zivilpersonen sterben, dann dürften die Entwickler das AWS gar nicht verkaufen, und sind sonst immer mitverantwortlich.
Das ist aber eine sehr starke eigene Voraussetzung, nach der man alle Waffenhersteller mitverantwortlich für die Toten macht inklusive

Küchenmesserhersteller, Alkoholhersteller, Pharmakonzerne etc., außer man grenzt das (un)gerechtfertigte Risiko näher ein.

ODER: Man könnte hier eine Unterscheidung aufmachen zwischen moralischer Verantwortung und Schuld und behaupten, dass sie moralisch verantwortlich waren, aber nicht schuld, weil das Risiko sehr gering war und das AWS von menschlichen Militärangehörigen überwacht werden sollte
→ Allerdings ist fragwürdig, wie es den Kern des Arguments widerlegen soll. Man müsste nur moralische Verantwortung durch Schuld bei allen beteiligten ersetzen und hätte dann ein Schuldzuschreibungsproblem

(b)

- P1a widerlegbar, aber trägt nicht aus wegen Oder in P1d; P1b schwierig zu widerlegen, P1d mit etwas Mühe vielleicht
- → Argument wird wahrscheinlich von den meisten akzeptiert, selbst wenn P1a nicht

3.

(a)

- **P2a:** Die Reaktionszeiten sind inzwischen zu kurz für den Menschen. (Überschallraketen brauchen 7s für 5km, Hyperschallraketen erreichen 5km in unter 1s.)
→ Reaktionszeiten sind offenbar sehr kurz bei Raketen und KNN's können sehr schnell sein, aber bei Soldaten mit Schusswaffen noch Zeit vorhanden, insbesondere wenn die Drohne klein, leise oder weit entfernt ist, und die Soldaten sind nicht kommen sehen; Use-Cases von AWS verschieden, für das Kriegsverbrechen oben gilt das nicht → min. Mitverantwortung von Offizieren; Allerdings könnten auch durch Raketen-Abwehr Kriegsverbrechen entstehen (z. B. Stadion mit Hunderten Zivilpersonen getroffen werden))
- **P2b:** Wenn die Reaktionszeiten inzwischen zu kurz für den Menschen sind, muss die Entscheidung zur Tötung vom AWS ohne den/die befehlshabende(n) Offizier(in) getroffen werden.
→ schwer zu widerlegen
- **P2c:** Wenn die Entscheidung zur Tötung vom AWS ohne den/die befehlshabende(n) Offizier(in) getroffen werden muss, kann dem/der befehlshabenden Offizier(in) nicht die moralische Verantwortung für Kriegsverbrechen, die von einem AWS unter ihrem Kommando begangen wurden, zugeschrieben werden.
- → noch schwerer zu widerlegen

(b) P2a kann widerlegt werden zumindest für viele Use-Cases; die Verantwortungszuschreibung bei autonomer Raketenabwehr bleibt aber ein Problem

4.

(a)

- **P3a:** Die AWS empfindet kein Leid.
→ Zweifel wären zumindest beim jetzigen Stand der Technik nicht überzeugend
- **P3b:** Wenn das AWS kein Leid empfindet, kann es nicht bestraft werden.
→ scheint recht überzeugend zu sein
- **P3c:** Wenn das AWS nicht bestraft werden kann, kann ihm nicht die moralische Verantwortung für Kriegsverbrechen, die vom ihm selbst begangen wurden, zugeschrieben werden.
→ analog zu P1d: Schuld von moralischer Verantwortung unterscheidbar; AWS könnte moralische Verantwortung tragen, weil es autonom agiert hat und kausal für das Kriegsverbrechen verantwortlich war, ohne aber schuld zu sein; aber dann Schuldzuschreibung statt Verantwortungszuschreibung problematisch: löst Problem nicht, verschiebt es nur

(b) wird wahrscheinlich akzeptiert

5. individuelle Lösung jeweils mit konkreten angegriffenen Prämissen basierend auf Lösung von 2. – 4.

- Option 1: eine der 2 Parteien ist verantwortlich/schuld
 - Entwickler
 - Offizier
- Option 2:
 - mehrere Parteien sind schuldig: Verantwortung wird je nach Situation aufgeteilt:
 - Entwickler, nur wenn Fehler nicht im vorher angegebenen Bereich
 - Offizier
- Option 3: eine andere Partei ist verantwortlich/schuld
 - z. B. Regierung, die AWS-Einsatz erlaubt/toleriert
- Option 4: niemand kann moralisch Verantwortung/Schuld zugeschrieben werden
 - AWS sollte wahrscheinlich nicht eingesetzt werden

Abbildungsverzeichnis

Hier werden nur Abbildungen aufgeführt, deren Quelle nicht offensichtlich ist. Screenshots als Vorschaubilder für Videos werden nicht aufgeführt. Der Übersichtlichkeit halber und für ein angenehmes Layout wurde verzichtet die Bilder im Deckblatt und die Bilder mit Saliency Maps aus dem Abschnitt Transparenz zu benennen.

Deckblatt

Obere Abbildung: <https://industrieweitschrift.de/wp-content/uploads/2025/01/autonome-bewaffnete-Konflikte-1054x602.jpg> (zuletzt abgerufen: 4.06.26)

Untere Abbildung: <https://tfppwire.com/wp-content/uploads/2025/02/l3harris-autonomous-swarm-solution-amorphous-hero.jpg> (zuletzt abgerufen: 4.06.26)

Abbildung 1: Diagramm A **Fehler! Textmarke nicht definiert.**

Screenshot aus TeachableMachine bei einem Beispiel-Bildprojekt Training/Details

Abbildung 2: Diagramm B **Fehler! Textmarke nicht definiert.**

Screenshot aus TeachableMachine bei einem Beispiel-Bildprojekt Training/Details

Abbildung 3: Fehlerberge und Täler **Fehler! Textmarke nicht definiert.**

https://www.isb.bayern.de/fileadmin/user_upload/Gymnasium/Faecher/Informatik/Handreichungen/Kuenstliche_Intelligenz_13/isb_kuenstliche_intelligenz_in_jgst_13.pdf. (S. 100)

Abbildung 4: Symbolbild für Interpretation der Situation durch AWS (KI-generiert) **Fehler! Textmarke nicht definiert.**

KI-generiert mithilfe von DALL E 3

Abbildung 5: Realität (KI-generiert, nicht erkannt per KI) .. **Fehler! Textmarke nicht definiert.**

KI-generiert mithilfe von DALL E 3

Abbildung 6: Soldaten ergeben sich einer Drohne – **Fehler! Textmarke nicht definiert.**

<https://www.merkur.de/assets/images/41/304/41304409-russische-soldaten-ergeben-sich-einer-landbasierten-drohne-NOBG.jpg> (zuletzt aufgerufen 31.07.26)

Abschnitt Transparenz

Eisert, Peter (2024): Deep Learning for Visual Computing, Vorlesung, Humboldt Universität. (Aufgabe 1-2)

Eisert, Peter (2024): Deep Learning for Visual Computing, Übung, Humboldt Universität, <https://drive.google.com/drive/folders/1PYTPHqJhav3jJvGGRMf0pU4gA8Td6Rsl>. (Aufgabe 3)

Literaturverzeichnis

Neutralität, Situationsverständnis und Verantwortung

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.

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

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.

Sparrow, Robert (2007): *Killer Robots*, in: *Journal of Applied Philosophy*, Vol. 24, No. 1, S. 66 – 67.

Ausblick

Altmann, Jürgen (2019): *Stellungnahme zur öffentlichen Sitzung des Unterausschusses Rüstungskontrolle, Nichtverbreitung und Rüstungskontrolle des Deutschen Bundestages am 6. November 2019*, Experimentelle Physik III, Technische Universität Dortmund.
<https://www.bundestag.de/resource/blob/687212/2a73a1d8af61f3bb2cc6f5b75232309c/Stellungnahme-Juergen-Altman-TU-Dortmund-data.pdf>

Burri, Susanne (2017): *What Is the Moral Problem with Killer Robots?*, in: Strawser, Jay (Hrsg.)/Jenkins, Ryan (Hrsg.)/Robillard, Michael (Hrsg.), *Who Should Die? The Ethics of Killing in War*, Oxford: Oxford University Press, S. 163 - 185.

Gadek, Guillaume (2024): *Unreliable AIs for the Military*, in: *Responsible Use of AI in Military Systems*, S. 101 – 123.

Madock, Jack (2025): *Robot warfare: the (im)permissibility of autonomous weapons systems*, in: *AI & Ethics*, 5(3), S. 2409 – 2417.

Renic, Neil/Schwarz, Elke (2023): *Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing*, in: *Ethics & International Affairs*, 37(3). S. 321–343.

ZDF (27.07.2023): *Autonome Kriegsmaschinen: Wo Drohnen mit Künstlicher Intelligenz im Einsatz sind*, <https://www.youtube.com/watch?v=XbKemARz8vo> [zuletzt abgerufen: 11.08.2025].