HAW HAMBURG

Der Skat-KI Turing-Test

Ermittlung der Eigenschaften glaubhafter Mitspieler-KI in Skat

Master-Thesis
zur Erlangung des akademischen Grades M.A.

im Studiengang Zeitabhängige Medien – Games

André Meyer [redacted]

Erstprüferin: Prof. Anke Günther

Zweitprüfer: Kolja Bopp

Hamburg, 21. 05. 2024

Hochschule für Angewandte Wissenschaften Hamburg
Fakultät Design, Medien und Information
Department Medientechnik

Inhaltsverzeichnis

1. Einleitung

Spiele sind seit einiger Zeit eine beliebte Testumgebung für die Forschung an künstlicher Intelligenz (KI). In Spielen wie Schach (Newborn, 1997) und Go (Silver et al., 2016, 2017) konnten bereits signifikante Erfolge verzeichnet werden. Dort spielen KIs schon seit Längerem besser als die stärksten menschlichen Spielenden.

Auch Videospiele sind davon nicht ausgenommen. In diesem Kontext werden KI-Systeme gemeinhin „Bots“ genannt. Für Echtzeitstrategie und verwandte Genres konnte deren Leistung bereits die menschlicher Spielender übertreffen (vgl. Starcraft 2 (Vinyals et al., 2019), Dota 2 (OpenAI et al., 2019)). Allerdings gilt dies nur in eingeschränkten Szenarien, mit weniger Komplexität als das vollständige Spiel.

In der Wissenschaft steht die KI vor allem auf der Gegenseite. Kooperation mit menschlichen Mitspielenden oder mit einer weiteren KI ist ein stark vernachlässigtes Thema (Ashktorab et al., 2020; Gao et al., 2022; Siu et al., 2021). Ein Beispiel für erfolgreiche Zusammenarbeit von Bots ist das Team von (OpenAI et al., 2019). In Dota 2 konnten sie innerhalb eines eingeschränkten Szenarios effektiv zusammen spielen. Kooperative KI bzw. Bots sind also gleichermaßen eine Herausforderung für Wissenschaft und Industrie.

Die Spieleindustrie ist sich bereits darüber einig, dass niemand gegen eine allmächtige KI antreten möchte. Ihr Ziel ist es dementsprechend, dem Menschen eine Illusion von Überlegenheit und knappen Siegen zu gewähren (Lidén, 2003; Lopez, 2005; West, 2009). Letztendlich möchte man gewinnen, jedoch nicht den Sieg geschenkt bekommen. Die perfekte KI verliere also absichtlich, ohne sich etwas anmerken zu lassen oder gar dumm zu wirken.

Eigentlich wünsche man sich einen anderen Menschen als Mitspielenden. Das belegt aktuelle Forschung. Es wurde zum Beispiel gezeigt, dass menschlich spielende Bots als unterhaltsamer empfunden werden (Baier et al., 2019; Cowling et al., 2015; Soni & Hingston, 2008). Zudem werden Bots, die zu maschinell wirken, von Menschen negativer wahrgenommen. Das gilt sowohl für ihre Kompetenz (Ashktorab et al., 2020) als auch für ihren Wert im Spielgeschehen (Merritt et al., 2011; Wehbe et al., 2017).

Bots für Skat müssen beide Disziplinen in sich vereinigen, da es sich um ein asymmetrisches Spiel für drei Spielende handelt. Somit wird Skat immer zwei gegen eins gespielt. Für den Kontext „Mensch gegen Computer“ lassen sich folgende Konstellationen ableiten:

– Zwei Bots gegen den Menschen,
– Ein Mensch zusammen mit einem Bot gegen den anderen Bot.

Ein Skat-Bot muss deshalb sowohl Kooperation als auch Konfrontation beherrschen, um erfolgreich zu spielen.

Ziel dieser Arbeit ist es, Glaubwürdigkeit und Kooperationsfähigkeit kommerzieller Skat-Bots zu untersuchen. Dafür soll eine Studie durchgeführt werden, die Erkenntnisse über menschliche oder eben maschinelle Spielweisen liefert.

Die Studie fokussiert sich dabei auf den kooperativen Teil von Skat, also auf das Team. Dadurch soll dazu beigetragen werden, die bestehende Forschungslücke anzugehen.

Erkenntnisse aus dieser Studie könnten in Zukunft dazu dienen, den Spielspaß und/oder die Kooperationsfähigkeit von Skat-Bots zu verbessern. In verschiedenen anderen Arbeiten haben diese Vorgehensweise bereits erfolgreich zur Verbesserung von Produkten genutzt (Baier et al., 2019; Ortega et al., 2013; Soni & Hingston, 2008).

1.1. Forschungsfragen

Aus der Zielsetzung ergeben sich folgende Forschungsfragen.

Wie glaubwürdig sind eingesetzte Bots?

Es stellt sich die Frage, ob Bots aus kommerziellen Titeln bereits für Menschen gehalten werden können. Die Glaubwürdigkeit der Bots wird im Rahmen eines Quasi-Turing-Tests ermittelt werden, ähnlich wie ihn verwandte Studien durchgeführt haben (Hingston, 2009; Laird & Duchi, 2000; Ortega et al., 2013). Grob gesagt sollen Studienteilnehmende, durch Beobachtung von einer Partie Skat, den gezeigten Spielenden als Mensch oder Bot identifizieren.

Sind Fähigkeit und Kooperation wichtig, um als Mensch zu gelten?

Die Spielstärke des Testsubjekts stellte in (Laird & Duchi, 2000) einen wichtigen Faktor für die Glaubwürdigkeit dar. Später konnte diese Beobachtung von (Hingston, 2009) nicht wiederholt werden. Ob ein Zusammenhang besteht, soll erneut untersucht werden.

Darüber hinaus hatte die Wahrnehmung als Maschine Einfluss auf die Kooperation (Ashktorab et al., 2020; Merritt et al., 2011; Wehbe et al., 2017). Insofern ist neben der Spielstärke zusätzlich interessant, ob als Menschen identifizierte Testsubjekte auch bessere Kooperation bescheinigt bekommen.

Warum werden Bots als solche identifiziert oder warum nicht?

Im Grunde genommen ist diese Frage eine Erweiterung der vorhergegangenen. Um ein besseres Verständnis darüber zu erlangen, wie Befragte ihr Urteil fällen, müssen sie es auch begründen. Auf diese Art und Weise soll festgestellt werden, welche Eigenschaften oder Züge mit Bots oder menschlichen Spielenden assoziiert werden.

Ähnlich wie in (Cowling et al., 2015; Milani et al., 2023) könnten Antworten auf diese Forschungsfrage zur Annäherung an menschliche Spielweise oder eben zur Vermeidung auffälliger Bot-Verhaltensweisen genutzt werden, um dadurch das Spielerlebnis zu verbessern.

Zudem könnten bestimmte demografische Eigenschaften der Testteilnehmenden hilfreich für die Aufgabe sein. Eigene Erfahrung mit Skat oder häufiges Spielen mit Skat-Bots sind vermutlich von Vorteil.

Welche Eigenschaften stehen für gute Kooperation?

Wie bei der Identifizierung soll auch die Bewertung der Kooperation nach Möglichkeit begründet werden. Hieraus könnten sich Verhaltensregeln ableiten lassen, die helfen Kooperation zu verbessern, analog zu (Correia et al., 2016).

2. Grundlagen und Verwandte Arbeit

Das Grundlagenkapitel dieser Arbeit beginnt mit einer grundlegenden Erklärung des Spiels Skat. Im Anschluss werden gängige KI-Methoden für Skat und deren bekannte Schwächen erklärt. Da die KI in Skat nicht nur gegen die Menschen spielt, sondern auch mit ihnen, werden ebenfalls Erkenntnisse aus der Forschung zu Mensch-KI-Kollaboration vorgestellt. Im dritten Unterkapitel wird erörtert, wie sich diese Erkenntnisse und der Konsens der unterhaltsameren menschlichen Bots auf das Design von kooperativen Bots auswirken. Abschließend wird mit dem Turing-Test für Bots ein Mittel vorgestellt, wie man Bots auf ihre Glaubwürdigkeit als Mensch testen kann.

2.1. Skat erklärt

Skat ist ein Kartenspiel für drei Spieler. Es handelt sich um ein Stichspiel, bei dem zwei Gegenspielende (GS) zusammen gegen einen Alleinspielenden (AS) antreten. Ein AS gewinnt die Runde, indem er mindestens 61 Punkte (Augen) in den Stichen gewinnt. Die Gegenpartei muss dies, ohne direkte Absprache, verhindern. In der Regel wird eine zuvor festgelegte Anzahl an Runden, die sog. Liste, gespielt. Wer am Ende die meisten Punkte holt, hat das Spiel gewonnen.

Ein Skat-Deck umfasst 32 Karten, welche aus jeweils acht Karten (Ass, Zehn, König, Dame, Bube, Neun, Acht, Sieben) der vier gängigen Farben (Kreuz, Pik, Herz, Karo) bestehen. Jede Karte hat einen bestimmten Wert, wobei die „Vollen" an der Spitze stehen. Diese sind das Ass (11 Punkte) und die Zehn (10 Punkte). Danach kommen König (4 Punkte), Dame (3 Punkte) und Bube (2 Punkte), gefolgt von den "Luschen" (Neun, Acht, Sieben), die keine Punkte wert sind. Insgesamt ergeben sich somit 120 verfügbare Punkte.

Zu Beginn jeder Runde werden an alle Spielenden 10 Karten ausgeteilt. Die zwei verbleibenden Karten sind der namensgebende Skat und werden bei Seite gelegt. Der Inhalt des Skats gehört dem AS. Während der Spielansage kann man mit ihm das eigene Blatt verbessern.

Die Spielansage ist dabei die zweite von den drei Phasen einer Runde, welche da sind:
- Reizen,
- Spielansage,
- Stichspiel.

Damit Zusammenhänge zwischen dem Stichspiel und den vorbereitenden Phasen Reizen bzw. Spielansage besser verstanden werden können, werden sie im Folgenden in umgekehrter Reihenfolge erklärt.

Das Stichspiel

Im Kern von Skat spielt jeder Spielende eine Karte aus. Diese drei Karten bilden einen Stich. Gespielt wird im Uhrzeigersinn, wobei die Reihenfolge von der Position der kartengebenden „Hinterhand" abhängt. Links von Hinterhand befindet sich „Vorhand", welche stets die erste Karte spielt. Danach kommt „Mittelhand" und dann Hinterhand selbst. In darauffolgenden Stichen beginnt dann, wer den vorherigen Stich gewonnen hat.

Innerhalb eines Stiches wird hierarchisch zwischen drei Arten von Karten unterschieden:
- Trümpfe,
- gültige Karten,
- ungültige Karten.

Trümpfe sind die vier Buben und alle Karten einer Farbe, die vorher vom AS während der Spielansage bestimmt wurde. Buben sind die stärksten Karten im Spiel, angefangen mit dem Kreuzbuben, dann Pik, dann Herz und schlussendlich Karo.

Eine Trumpfkarte schlägt jede nicht-trumpf Karte. Die Hierarchie der anderen Karten ist Ass, Zehn, König, Dame, Neun, Acht, Sieben.

Für das Ausspielen von Karten gilt allgemein die „Bedienregel“. Die erste Karte gibt die zu bedienende Farbe vor, alle folgenden Karten müssen also die gleiche Farbe haben. Sie sind gültige Karten. Sollte diese Regel nicht erfüllt werden können, kann eine beliebige Farbe abgeworfen oder Trumpf gespielt werden.

Der höchste ausgespielte Trumpf gewinnt den Stich. Sollte kein Trumpf ausgespielt worden sein, gewinnt die höchste gültige Karte. Abgeworfene und damit ungültige Karten können zwar keinen Stich gewinnen, sie sind jedoch weiterhin Punkte wert. Gegenspielende werfen gerne Karten mit hohem Wert ab, wenn ersichtlich ist, dass der Stich an Teammitglieder geht. Ein solcher Spielzug wird als „Schmieren“ bezeichnet. Eine gängige Strategie des AS besteht darin, die GS dazu zu zwingen, Trümpfe zu bedienen. Dadurch kann ein Trumpfmonopol erlangt werden, weil in der Theorie jeder eröffnenden Trumpfkarte des AS durch die Bedienregel zwei weitere Trumpfkarten der GS folgen müssen.

Insgesamt werden so zehn Stiche gespielt. Nach dem letzten Stich werden die Punkte in den gewonnenen Stichen beider Parteien gezählt. Die Punkte im Skat zählen dabei zu denen des AS. Sobald eine Partei über die Hälfte aller Punkte gesammelt hat, gilt die Runde als gewonnen. Im Falle eines Sieges der GS, werden dem AS Listenpunkte abgezogen. Dies ist auch bei Unentschieden der Fall. Gewinnt der AS, werden ihm Listenpunkte gutgeschrieben.

Über die für den Sieg erforderlichen 61 Punkte hinaus kann es für den AS notwendig sein, weitere Bedingungen zu erfüllen. Diese ergeben sich aus der getroffenen Spielansage, auf die im Folgenden eingegangen wird.

Die Spielansage

In der Spielansage muss der AS verschiedene Entscheidungen treffen, welche das Stichspiel sowie die Siegbedingungen beeinflussen. Dazu zählen:

– Welches Spiel wird angesagt?
– Wird der Skat aufgenommen oder nicht?
– Verpflichtet man sich dazu, höhere Gewinnstufen zu erreichen?

Allem voran gibt das angesagte Spiel die Trumpffarbe vor. Das gilt jedenfalls für die vier sog. „Farbspiele“ Kreuz, Pik, Herz, Karo. Im ersten Spezialspiel „Grand“ sind ausschließlich die Buben Trumpf. Gar keine Trümpfe gibt es dann im zweiten Spezialspiel „Null“ oder „Nullspiel“.

Innerhalb eines Nullspiels versucht der AS keinen einzigen Stich zu gewinnen. Sollte dieser Fall trotzdem eintreten, ist die Runde sofort verloren. Die Gegenpartei will deshalb bewusst ihre Stiche verlieren. Darüber hinaus ist in diesem Spiel die Hierarchie der Karten verändert. Die neue Reihenfolge lautet wie folgt: Ass, König, Dame, Bube, Zehn, Neun, Acht, Sieben.

Noch bevor der AS sich allerdings auf ein Spiel festlegt, besteht optional die Möglichkeit, den Skat aufzunehmen. Im Normalfall wird dies auch getan, da mit den Karten aus dem Skat das eigene Blatt verbessert werden kann, indem man sie mit (schlechteren) Handkarten austauscht. Dieser Vorgang wird „Drücken“ genannt. Abschließend werden die zwei gedrückten Karten verdeckt zurück auf den Tisch gelegt. Wird der Skat hingegen unberührt gelassen, spricht man von einem „Handspiel“.

Ein Handspiel birgt ein erhöhtes Risiko, da weder das Blatt aufgebessert noch die Karten aus dem Skat angeschaut werden dürfen. Es stellt die erste zusätzliche Gewinnstufe dar, die angesagt werden kann. Ferner noch ermöglicht es das Ansagen der weiteren Gewinnstufen, die da sind: „Schneider“, (Schneider-) „Schwarz“ und „Ouvert“.

Die Gewinnstufen sind eine lineare Skala. Somit wird auch Schneider mitangesagt, wenn Schwarz angesagt wird, und jeweils beide, wenn Ouvert angesagt wird.

Mit diesen Ansagen verpflichtet sich der AS, mindestens 90 Augen zu holen (Schneider), alle Stiche zu gewinnen (Schwarz) und/oder die Handkarten offen auf den Tisch zu legen (Ouvert). Bei einem Nullspiel sind nur die Zusätze Hand und Ouvert möglich, da keine Stiche gewonnen werden sollen. Demnach bedeutet die Ansage Null Ouvert lediglich, dass man mit offenen Handkarten spielt. Um das Wegfallen der Gewinnstufen Schneider und Schwarz auszugleichen, kann Null auch weiterhin offen gespielt werden, wenn bereits der Skat aufgenommen wurde.

Der Reiz daran, höhere Gewinnstufen anzusagen, liegt für den AS darin, dass jede Gewinnstufe einen Multiplikator erhöht, welcher den Wert des Spiels und die daraus resultierenden Listenpunkte steigen lässt. Jede Gewinnstufe erhöht den Multiplikator um eins.

Für Spiele ohne weitere Ansagen gelten die Grundwerte der Spiele:
- 24 für Grand,
- 12 für Kreuz,
- 11 für Pik,
- 10 für Herz,
- 9 für Karo.

Davon ausgenommen sind die vier verschiedenen Ansagen von Nullspielen. Sie verfügen über feste Werte (z.B. 23 für normales Null).

Ergänzend ist noch zu sagen, dass die Gewinnstufen Schneider und Schwarz auch ohne Ansage erreicht werden können. Dies ist der Fall, wenn sich am Ende einer Runde herausstellt, dass 90 oder mehr Punkte gewonnen wurden oder gar alle Stiche. Man spricht dann von einem stillen Schneider. Dementsprechend kann auch Schneider bzw. Schwarz von der Gegenpartei gewonnen werden, wodurch der AS noch mehr Listenpunkte verliert.

Als Ausgleich dafür erhöht das konkrete und damit verpflichtende Ansagen von Schneider oder Schwarz den Multiplikator jeweils um zwei. Somit können die zusätzlichen Gewinnstufen den Multiplikator um insgesamt sechs erhöhen.

Die Spielenden planen bereits nach dem Verteilen der Karten ihre Ansage, da die Wertigkeit des angestrebten Spiels gleichzeitig auch die Stärke während des Reizens beeinflusst. Wie das Reizen abläuft, soll als nächstes erklärt werden.

Das Reizen

Das Reizen ist eine Auktion, bei der die Rolle des AS zu ersteigern ist. Die Höhe des Gebots („Reizwert“, „Reizgebot“) ist dabei direkt abhängig vom eigenen Blatt. Maßgeblich für die Reizwerte ist der sog. „Spitzenwert“. Dieser gibt die Anzahl an lückenlos aufeinanderfolgenden Buben bzw. in Reihe fehlenden Trümpfen an.

Für die Berechnung des Spitzenwerts wird zunächst unterschieden, ob sich der Kreuzbube auf der Hand befindet:

- **Mit Kreuzbube**
  Gezählt werden lückenlos aufeinanderfolgende Buben. Nur Kreuzbube auf der Hand ist „mit eins“, Kreuz und Pikbube „mit zwei“, bis hin zu „mit vier“ (alle Buben auf der Hand). Beispiel für eine Lücke wäre Kreuz- und Herzbube, wo der Pikbube fehlt, weswegen nach eins aufgehört wird zu zählen.
- **Ohne Kreuzbube**
  Gezählt werden dem Rang nach fehlende Trumpfkarten, bis zum ersten vorhandenen Trumpf. Wenn beispielsweise nur der Pikbube auf der Hand ist, spricht man von „ohne eins“. Fehlt auch der Pikbube, aber der Herzbube ist vorhanden, ist man „ohne zwei“. Eine Hand ohne Buben, aber mit Trumpf-Ass ist „ohne vier“ usw. bis hin zu „ohne elf“, was gar keine Trümpfe in der angesagten Farbe bedeutet.

Auf den Spitzenwert wird dann der zuvor erwähnte Multiplikator addiert, welcher gegebenenfalls durch die Ansage weiterer Gewinnstufen entstanden ist. Für das Wagnis, AS zu werden, ist der Grundwert für den Multiplikator gleich eins, ohne dass es irgendeiner Art von zusätzlichen Ansagen bedarf. Die Summe der beiden Werte wird final mit dem Grundwert des angesagten Spiels addiert. Letztendlich ergibt sich so eine diskrete Skala mit Werten zwischen 18 (z.B. Karo mit einem Buben; 9x2) und 264 (Grand Ouvert mit vier Buben; 24x11) als möglichen Geboten.

Geboten wird reihum, wobei die Reihenfolge wieder von Hinterhand abhängt. Es gilt der Merksatz „Geben, Hören, Sagen“. Hinterhand hat die Karten ausgegeben und Vorhand hört das erste Gebot, das von Mittelhand ausgesprochen wird.

Als Sagender muss man sich immer entscheiden, ob man das nächsthöhere Gebot sagt oder passt. Hörende bestätigen Gebote oder passen ihrerseits. Es können nur bestimmte Rollen eingenommen werden, die von der eigenen Position abhängen.

Vorhand darf nur hören, Mittelhand kann hören und sagen, Hinterhand darf nur sagen. Entsprechend werden die Rollen gewechselt, wenn Vorhand gegen Mittelhand passt. Hinterhand sagt nun die Gebote an, während Mittelhand hört, da Hinterhand selbst nicht hören darf.

Sollten alle Spielenden sofort passen, gilt die Runde als „eingepasst“. Es wird neu gestartet. Falls nur zwei Spielende passen, hat der dritte das Reizen gewonnen und wird somit AS. Dieser Spielende führt dann die Spielansage durch.

Die mit den Reizwerten einhergehende Wertigkeit eines Spiels muss eingehalten werden. Konkret bedeutet dies, dass die angesagte Spielstärke nicht unterschritten werden sollte. Diese Regel wird allerdings erst bei der Spielauswertung überprüft. Beispielsweise sollte ein AS mit Reizgebot 20 nicht die geringere Ansage „Karo mit einem“ tätigen. Stellt sich bei der Spielauswertung heraus, dass zu viel geboten wurde, hat man „überreizt“ und damit verloren. Das Überreizen kann auch durch das Finden von Buben im Skat herbeigeführt werden oder durch das Erreichen eines stillen Schneiders vermieden werden. Aus diesem Grund erfolgt die Kontrolle erst am Ende des Spiels.

Weil Gebote eingehalten werden müssen und sie direkt mit den Handkarten zusammenhängen, teilen sie gleichzeitig auch viele Informationen mit dem Tisch. Ein Spielender, der auf 24 reizt, möchte vermutlich Kreuz spielen. Er wird viele Kreuz-Karten und garantiert einen Buben besitzen. Wenn ein Spielender bis 23 mitgeht, wurde ein Nullspiel angestrebt und seine Hand ist entsprechend schlecht für andere Spiele.

Nachdem Skat nun grundlegend erklärt wurde, geht es im nächsten Kapitel darum, wie KIs das Spiel spielen.

2.2. KI für Skat

In dem Buch „AI Methods“ beschreiben Yannakakis und Togelius gängige Methoden zur Implementierung von KI in Spielen. Speziell für Skat relevante Methoden daraus sind (Yannakakis & Togelius, 2018):

– Regelbasierte Systeme („Ad-Hoc Behavior Authoring“),
– Suchbäume („Tree Search“),
– Maschinelles Lernen („Supervised Learning“ und Fortfolgende).

Regelbasierte Systeme evaluieren den Spielstand basierend auf handgeschriebenen Heuristiken, die meist auf menschlichem (Experten-)Wissen basieren. Im Grunde genommen kann man sie als Verkettungen von „wenn-dann-sonst“-Blöcken sehen (Niklaus et al., 2019).

Für Skat könnten einfache Regeln beispielsweise sein:

– Wenn du Herz spielen willst, drücke niemals Herzkarten,
– Folge der Farbe deines Mitspielenden,
– Bringe den AS in Mittelhand.

Eine regelbasierte KI für Skat ist „XSkat“ von Gunter Gerhardt (Gerhardt, 2004). Da die Software frei verfügbar ist, wurde sie von einigen Arbeiten als Referenz genutzt (Buro et al., 2009; Kupferschmid, 2003; J. Schäfer et al., 2008).

Der grundlegende Vorteil von XSkat und regelbasierten Systemen im Allgemeinen liegt in der hohen Geschwindigkeit, mit der diese ihre Züge berechnen können. Aufgrund des statischen Aufbaus ist der Rechenaufwand gering. Allerdings führt diese Statik zu einer hohen Vorhersehbarkeit und somit zu einer geringen Spielstärke, wie auch in den zuvor genannten Arbeiten beschrieben wird.

Eine tatsächlich spielstarke Mischung aus regelbasiert und Suchbäumen ist die Fox-KI von (Edelkamp, 2021a). Ihre Heuristiken nutzen eine umfangreiche Datenbank von Expertenspielen, die zusammen mit dem aktuellen Spielstand zum Ableiten von Entscheidungen herangezogen wird. Im Detail werden die Entscheidungen mithilfe der mathematischen Methoden von (Gößl, 2019) getroffen.

Die Fox-KI erreichte auf diese Weise menschliches Expertenniveau. Insbesondere wurde die Spielstärke im Nullspiel erhöht, welches für die meisten kommerziellen Skat-KIs einen Schwachpunkt darstellt (Gößl, 2020).

Wie bereits erwähnt wurden in der Fox-KI auch Suchbäume genutzt. Im weitesten Sinne kann ein Suchbaum als eine Verkettung von Knoten definiert werden, die jeweils Zustände (im Spiel) repräsentieren. Diese Zustände sind dann über Aktionen bzw. Äste miteinander verbunden (Yannakakis & Togelius, 2018). Normalerweise gibt es an jedem Knoten, der nicht das Ende des Spiels darstellt, mehrere ausführbare Aktionen, die dann wieder zu neuen Knoten führen.

Entscheidend für die Verwendung der Bäume und die Unterschiede zwischen Algorithmen ist, welche Äste in welcher Reihenfolge betrachtet oder gar ignoriert werden (Yannakakis & Togelius, 2018).

Ein nennenswerter Algorithmus ist die „Montecarlo-Methode“ (Montecarlo Tree Search, MCTS) genutzt. Der Ansatz basiert auf der Idee, mit Hilfe von zufälligen Stichproben eine Vielzahl von Aktionen eines (zu) großen Aktionsraumes zu evaluieren, um in angemessener Zeit den bestmöglichen Zug zu ermitteln (Yannakakis & Togelius, 2018).

Skat ist jedoch ein Spiel mit verdeckten Informationen. Die Spielenden kennen anfangs ausschließlich die eigenen Handkarten. Weitere Karten oder der Inhalt des Skats werden erst im Verlauf des Spiels offengelegt. Dies führt dazu, dass in Schach oder Go erfolgreiche

Suchalgorithmen, wie MCTS, in Skat nicht ohne weiteres funktionieren (Ginsberg, 1999; Kupferschmid, 2003).

Aus diesem Grund wurde der Algorithmus von (Ginsberg, 1999) für seine Bridge KI „GIB“ abgeändert. Der so entstandene „Perfekte-Informationen-Montecarlo-Algorithmus“ (Perfect Information Montecarlo, PIMC) wurde dann zum ersten Mal von (Kupferschmid, 2003) für Skat angewendet.

Kurz gesagt funktioniert der Algorithmus wie folgt:

- Konstruiere eine Menge von Kartenverteilungen, die konsistent mit den bereits gesammelten Informationen ist,
- Simuliere für jede Kartenverteilung und spielbare Karte, den weiteren Spielverlauf mit offenen Karten,
- Spiele die Karte, die die meisten Spiele gewonnen hat.
- Haben mehrere Karten Gleichstand, dann spiele die Karte, die in ihren Spielen aufsummiert die höchste Punktzahl hat.

Die Performance dieses Ansatzes ist abhängig von der Anzahl an Simulationen („Welten“), die berechnet werden dürfen. Bereits eine Reduktion von 100 Welten auf 50 ließ GIB signifikant schlechter spielen. Mit einer geringeren Anzahl an Welten steige schlichtweg die Wahrscheinlichkeit, dass der beste Zug gar nicht in diesen vorkommt und somit nicht gefunden werden könne (Ginsberg, 1999; Kupferschmid & Helmert, 2007).

Weitere Nachteile ergeben sich durch den Wechsel in ein offenes Spiel (Frank & Basin, 1998; Long, 2011). Beispielsweise bringt der Algorithmus dadurch nie Züge hervor, die dem reinen Informationserwerb dienen (Ginsberg, 1999; Kupferschmid, 2003). Da er im Glauben ist, alle Karten zu kennen, lohnen sich solche Züge nicht, weil sie meist eingesetzte Karten verlieren.

Im Gegensatz dazu würden Menschen den Verlust einer Karte möglicherweise in Kauf nehmen, da sie auf diese Weise neue Informationen über die Karten ihres Gegenübers erhalten können. Wenn eine Farbe noch nie ausgespielt wurde, kann dort eine Lücke im gegnerischen Blatt vorliegen. Ob dies der Fall ist, lässt sich jedoch nur durch Anspielen der Farbe feststellen.

PIMC kann zudem fürs Reizen und das Drücken des Skats genutzt werden, indem aus Geboten bzw. gedrückten Karten gebildete Welten auf die gleiche Art und Weise simuliert werden (Keller & Kupferschmid, 2008; Kupferschmid, 2003).

Bei der Evaluierung potenzieller Spielansagen können gewisse Kombinationen, mit perfekten Informationen, unspielbar erscheinen, obgleich sie dies in der Realität nicht sind. Wenn man beispielsweise nach dem Drücken mit einer einzelnen hohen Karte auf der Hand verbleibt, scheint es unmöglich, ein Nullspiel zu gewinnen. Gegenspielende mit perfekten Informationen würden diese Schwäche sofort anspielen und das Spiel beenden. In der Realität haben sie aber keine Ahnung, dass diese Schwäche in der Hand des AS existiert. Somit wird die Möglichkeit ein Nullspiel anzusagen unnötigerweise verworfen.

Dementsprechend sind die Schwächen der meisten KIs ein zu sicheres bzw. konservatives Verhalten vor dem Stichspiel (Buro et al., 2009; Kupferschmid, 2003; J. Schäfer et al., 2008). Es werden keine Serien gegen Menschen gewonnen, weil häufig weniger (wertvolle) Spiele angesagt werden. Aus diesem Grund verfolgen diverse Arbeiten andere Ansätze, um den Spielablauf vor dem Stichspiel zu optimieren (Edelkamp, 2021b; Keller & Kupferschmid, 2008; Rebstock et al., 2019).

Für das Stichspiel selbst sind Suchbäume bzw. PIMC-Ansätze bis heute die stärkste Methode (Long, 2011; Rebstock et al., 2019). So wurde das Niveau von menschlichen Experten erreicht, aber noch nicht übertroffen (Buro et al., 2009; Edelkamp, 2021a).

Um die Spielstärke noch weiter zu erhöhen, wird am häufigsten daran gearbeitet, die Evaluation von Welten und einzelnen Zügen zu verbessern (Buro et al., 2009; Edelkamp, 2020; Long, 2011).

Eine entsprechende Erweiterung der Arbeiten von Kupferschmied et al. stellt die KI „Kermit“ von (Buro et al., 2009) dar. Diese wurde in den nachfolgenden Jahren stetig verbessert und ist momentan womöglich die stärkste KI (ein Vergleich zwischen Kermit und Fox fehlt aufgrund mangelnder Kompatibilität (Edelkamp, 2021a)).

Kermits Stichspiel wurde verbessert, indem mehr Schlussfolgerungen aus Reizwerten und bisher gespielten Karten gezogen werden (Buro et al., 2009). Die Spielweise der Gegenseite wird ebenfalls von Kermit analysiert (Long, 2011). Aufgrund der verbesserten Informationslage können somit realistischere Welten bevorzugt evaluiert werden und unrealistische Welten schneller verworfen werden.

Als letzte Optimierung wird maschinelles Lernen bei Kermit eingesetzt. Algorithmen dieser Art „lernen“ Modelle oder „trainieren“ Netzwerke basierend auf großen Datenmengen, die markierte Eigenschaften für die Lernaufgabe enthalten (Yannakakis & Togelius, 2018).

Um Kermits Reizen und Spielansage zu verbessern wurde ein Netzwerk mit Daten aus menschlichen Partien trainiert (Rebstock et al., 2019). Indem die KI menschlichen Spielstil imitiert, konnten Reizen, Spielansage und Drücken signifikant verbessert werden.

Einzig das Stichspiel war noch immer schwächer als die PIMC-Methode, da diese gerade im späteren Spielverlauf perfekt spielt. Daher wurde die Entwicklung eines weiteren Modells beschlossen, welches bei der Auswahl von Welten hilft (Solinas et al., 2019). Auf diese Weise wurde dann auch das Stichspiel stärker als die vorherige Version.

Abschließend kann man bei den erprobten Ansätzen für maschinelles Lernen und Skat vor allem sehen, dass diese wesentlich schneller als normale Suche sind (Rebstock et al., 2019; Solinas et al., 2019). Dieser Aspekt ist wesentlich für das Spielen mit Menschen in Echtzeit. Kürzeres Warten auf die KI kann dort eine flüssigere Spielerfahrung bewirken. Mit PIMC ist dies nicht möglich, da zur Verfügung stehende Zeit maßgeblich für deren Spielstärke ist.

Wie eingangs bereits erwähnt liegt der Fokus von KI oft darin, ein interessanter Ersatz für menschliche Mitspielende zu sein. Skat-KIs treten jedoch gleichzeitig als Teammitglieder auf. Entsprechend gibt es auch Erkenntnisse der Autoren über Kermits Fähigkeit zur Kooperation, die im nächsten Kapitel als Einstieg in das Thema „Mensch-KI-Kollaboration“ dienen sollen.

2.3. Mensch-KI-Kollaboration

Im vorherigen Kapitel wurde bereits dargelegt, dass die Annahme, über perfekte Informationen zu verfügen, zu Fehlern führt und das Gewinnen von Informationen vernachlässigt wird. Ebenso kann Kommunikation mit Teammitgliedern ausbleiben, weil davon ausgegangen wird, dass alle Spielenden perfekt spielen können.

In (Long, 2011) erläutert der Autor, dass Kermits Züge für seine Mitspieler entsprechend ambivalent sein können:

„Angenommen die letzte Trumpfkarte sei der Kreuzbube. Der AS hält diesen und eine Pik-Sieben in der Hand. GS1 hat Pik-Acht und Neun, GS2 hat Herz-Ass und Sieben. Wer das Ass gewinnt, gewinnt das Spiel. GS1 kommt raus, dann GS2, dann der AS. Aus der Sicht perfekter Information ist für GS1 irrelevant, welche Karte gespielt wird. Beide seiner Karten schlagen die Pik-Karte des AS. Sein Teammitglied muss einfach das Ass schmieren.

Angenommen GS2 kennt ebenfalls die verbliebenen Karten, weiß aber nicht, wie diese verteilt sind. Wenn GS1 jetzt die Pik-Acht spielt, weil es aus seiner Sicht egal ist, steht GS2 vor einem unlösbaren Dilemma.

In der beschriebenen Version muss sofort das Ass gespielt werden, aber aus Sicht von GS2 könnte der Bube genau so gut bei GS1 und nicht beim AS liegen. Für diesen Fall müsste GS2 das Ass halten und im letzten Stich spielen, damit sie gewinnen. Spielt GS2 das Ass auf die Pik-Acht könnten sie verlieren.

Behoben werden kann dieses Kommunikationsproblem nur, wenn GS1 seine gewinnende Karte – Pik-Neun oder Kreuzbube, welche auch immer er hat – zuerst spielt."

Man sieht, dass Kommunikation zwischen den Teammitgliedern über die Karten erfolgt. Das Spielen der höchsten Karte signalisiert dem Mitspielenden, sein Ass zu schmieren. Aus diesem Grund haben sich in Skat verschiedene Regeln etabliert, wie man auf die Karten seines Mitspielenden reagieren sollte. Beispielsweise ist es geboten, Farben nachzuspielen, die ein Teammitglied zuvor aufgespielt hat.

Diese Regeln sind KIs nicht unbedingt bekannt, was die Kommunikation mit menschlichen Mitspielenden erschwert. Insbesondere KIs, die Spiele durch das Spielen gegen sich selbst erlernt haben, neigen dazu, eigene Strategien zu entwickeln, die anders als menschliche sind (Rebstock et al., 2019; Siu et al., 2021).

Entsprechend äußerten Studienteilnehmende in (Siu et al., 2021) eine Präferenz für einen regelbasierten Bot, gegenüber einem durch Selbstspiel trainierten Bot. Sie konnten sich eher auf die Spielweise des regelbasierten Bots einstellen, da sie ihn verstanden. Der Bot mit maschinellem Lernen hingegen war auf eine andere Grundstrategie als menschliche Spielende trainiert worden und wies zudem neuartige Konventionen auf.

Man kann erkennen, dass Menschen ein mit der KI geteiltes Verständnis der Aufgabe erwarten. Dies wird ebenfalls in (Zhang et al., 2021) festgestellt. Die Dissonanz zwischen ihrem Grundverständnis des Spiels und dem des Bots führte dazu, dass Spielende schnell weniger Vertrauen in den Selbstspiel-Bot hatten, obwohl die Spielresultate mit beiden Bots gleich gut waren.

Vertrauen ist ein wichtiger Aspekt für Kooperation, der in (Correia et al., 2016) vergleichend zwischen menschlichen und KI-Mitspielenden betrachtet wurde. Die Ergebnisse zeigen, dass das Vertrauen in Mitspielende von der Erfahrung im Zusammenspiel abhängt. Das anfängliche Vertrauen in die KI war dabei geringer als in zuvor unbekannte Menschen. Es stieg dann aber durch wiederholtes Spielen signifikant an, während es bei menschlichen Partnern konstant war.

Das zunächst geringere Vertrauen lässt darauf schließen, dass es eine Art Misstrauen gegenüber KI-Teammitgliedern gibt. Ein möglicher Grund dafür ist die wahrgenommene oder vermutete geringere Kompetenz der KI (Ashktorab et al., 2020; Correia et al., 2016; Merritt et al., 2011; Siu et al., 2021; Wehbe et al., 2017). Im Experiment hat die KI gut gespielt, weswegen das Vertrauen in sie stieg.

Tatsächlich wird mechanische Kompetenz der KI von menschlichen Teammitgliedern als am wertvollsten empfunden, noch vor geteiltem Verständnis (Zhang et al., 2021).

Die beobachteten Umstände legen nahe, dass KI von ihren Mitspielenden als Werkzeug wahrgenommen wird und nicht als gleichwertiges Teammitglied. Im Gegensatz zur KI als Gegenspielerin hat sie keine eigene Agenda im Spiel. Ihr Zweck besteht ausschließlich darin, den Menschen beim Gewinnen zu unterstützen.

Diese Annahme wird von den Erkenntnissen in (Merritt et al., 2011) untermauert. In ihrer Studie betrachteten sie Spielende während eines kooperativen Spiels. Dabei wurde behauptet, dass ihr Teammitglied entweder eine KI oder ein anderer Mensch sei, was in einigen Fällen falsch war.

Wenn Spielende glaubten, mit anderen Menschen zusammenzuspielen, waren sie vorsichtiger darin, diese für ihre Fehler zu beschuldigen. Glaubten sie hingegen, mit einer KI zu spielen, wurde diese häufig für das Versagen des Teams verantwortlich gemacht.

Unterschiedliches Verhalten gegenüber Bots wurde ebenfalls in (Wehbe et al., 2017) und (Ashktorab et al., 2020) festgestellt.

In Wehbe et al. spielten Testteilnehmende das kooperative Spiel Left 4 Dead. Ihnen war dabei nicht klar, ob sie mit Menschen oder Bots zusammenspielten. Im Experiment stellte sich eine Hierarchie zwischen Mensch und Bot heraus. Bots werden als Beifahrende definiert, die menschliche Stars bei ihren Handlungen unterstützen. Wann immer ein Bot tatsächlich gute, proaktive Entscheidungen traf oder Können bewies, wurde er als menschlicher eingestuft.

Es wird also keine Initiative von Bots erwartet. Der Mensch führt, der Bot folgt. Er soll sich auf den Menschen einstellen. Genau diese Anpassungsfähigkeit sei jedoch eine menschliche Kompetenz, die Bots meistens nicht besäßen. So jedenfalls die Testenden in (Ashktorab et al., 2020).

Abschließend kann man festhalten, dass sich das Design von Bots für bessere Kooperation verändern muss. Inwiefern ein menschlicherer Spielstil dabei helfen könnte und warum gängige Bot-Designprinzipien überhaupt zur Einordnung als Werkzeug geführt haben, beschreibt das nächste Kapitel.

2.4. Designprinzipien für kooperative Bots

Es ist unter Spielenden weithin bekannt, dass Bots in der Lage sind, mechanisch perfekt zu spielen. Ferner wird genau dies von kooperativen Bots gefordert (Correia et al., 2016; Zhang et al., 2021). Da ein fehlerloser Bot kaum zu schlagen wäre, muss er den Menschen den Sieg eigenhändig ermöglichen.

Aus diesem Umstand heraus sind Schlagwörter wie „Künstliche Dummheit“ und „Intelligente Fehler“ entstanden (Lidén, 2003; Lopez, 2005; West, 2009). Mit ihnen ist gemeint, den Menschen gewinnen zu lassen, ohne den Bot dabei dumm aussehen zu lassen. Ferner soll der Mensch ein Gefühl von Intelligenz und Überlegenheit vermittelt bekommen. Allerdings darf das Ganze nicht auffallen, denn niemand lässt sich den Sieg gerne schenken.

(Lopez, 2005) beschreibt in diesem Sinne, wie Schach-Bots eine positive Spielerfahrung für unterschiedliche Niveaus erzeugen, indem sie subtile Fehler machen, um dem menschlichen Gegenüber einen oder mehrere Vorteile zu gewähren. Sobald der Vorteil angenommen wurde, spielt die KI ungehindert weiter, was hoffentlich zu einem knappen Sieg für den Menschen führen sollte.

(Lidén, 2003) empfiehlt beispielsweise, Bots in Shootern verringerte Genauigkeit zu geben oder sie stets den ersten Schuss verfehlen zu lassen. Das Verfehlen von Schüssen vermittle den Spielenden das Gefühl, sich clever bewegt zu haben. Gerade knappe Fehlschüsse aus unbekannter Richtung erzeugen dabei viel Spannung.

Es ist erkennbar, dass das Design die Spielenden in den Mittelpunkt stellt. Die Bots halten sich zurück, um ihnen eine bessere Erfahrung zu bieten. Für Kooperation heißt das, den menschlichen Mitspielenden zur Seite zu stehen und ihnen Möglichkeiten für großartige Züge zu bieten. Selbst wenn ein Bot den besten Zug erkannt hat, sollte er ihn nicht unbedingt machen (West, 2009), weil er den Spielenden vorbehalten ist.

Genau diese Philosophie führt zur Wahrnehmung von Bots als nützliche Statisten, die in (Wehbe et al., 2017) beschrieben wird. Da sie keiner eigenen Agenda folgen, scheint ihr Mitwirken für den Spielausgang unwichtig zu sein (Ashktorab et al., 2020).

Vielleicht auch deswegen beschreiben Befragte in (Ashktorab et al., 2020) eine qualitativ schlechtere Erfahrung, wenn sie der Meinung sind, mit einem Bot zu interagieren.

Ein Ansatz zur Verbesserung der Spielerfahrung mit Bots als Gegenspielern wurde von (Soni & Hingston, 2008) erforscht. Hierzu trainierten sie neuronale Netze mit menschlichen Testdaten, um Bots mit menschenähnlichem Spielstil zu erzeugen. Anschließend wurden diese mit einem regelbasierten Bot verglichen. Den Befragten der Studie zu Folge lieferten die trainierten Netze qualitativ bessere Bots. Genauer gesagt waren sie weniger vorhersehbar, herausfordernder, menschenähnlicher und hatten höheren Wiederspielwert.

Aus den Erkenntnissen von Hingston und Ashktorab kann man ableiten, dass sich auch kooperative Bots durch ein menschenähnliches Verhalten verbessern lassen.

Das Training eines Bots mit menschlichen Daten fördert das geteilte Verständnis von Mensch und Bot. Wenn der Bot bereits wie der Mensch spielt, muss sich der Mensch nicht auf seinen (neuartigen) Spielstil einstellen, weil sich der Bot bereits auf ihn eingestellt hat. Es besteht die Möglichkeit, dass Menschen die Züge des Bots besser verstehen, da sie selbst genau so spielen würden. Besseres Verstehen verstärkt dann das Vertrauen in den Bot. All dies sind Dinge, die Quellen und Befragte aus dem vorherigen Kapitel gefordert haben (Ashktorab et al., 2020; Siu et al., 2021; Zhang et al., 2021).

Forschung, die einen solchen Ansatz verfolgt, gibt es beispielsweise zu dem Kartenspiel Spades.

Durch Analyse von Spieldaten der marktführenden Spades-App wurde festgestellt, dass sich die Züge des verwendeten MCTS-Bots signifikant von denen der menschlichen Teammitglieder unterscheiden (Cowling et al., 2015).

Aufgrund dieser Erkenntnis wurde ein neuronales Netz mit den erhobenen Daten trainiert, um die Unterschiede von menschlichen Zügen zu denen des Bots zu verringern (Baier et al., 2019; Devlin et al., 2016). Mit dieser Vorgehensweise konnte durch direkte Imitation ein Bot erstellt werden, der bei gleichbleibender Spielstärke den menschlichen Spielstil emuliert. Ähnliche Erfolge erzielten auch (Khalifa et al., 2016), von deren Arbeit der Spades-Bot inspiriert ist.

Eine empirische Bewertung der menschlichen Spielweise des Spades-Bots wurde zwar von den Forschenden vorgeschlagen, bislang jedoch nicht durchgeführt. Für eine solche Art von Test wird in aktueller Forschung eine Abwandlung des bekannten Turing-Tests verwendet. Innerhalb der Tests wird ermittelt, wie glaubwürdig KIs bzw. Bots als Mensch wirken. Die Struktur eines solchen Tests wird im nachfolgenden Kapitel erläutert.

2.5. Turing-Tests für Bots

Der originale Turing Test wurde von Alan Turing im Jahr 1950 beschrieben. Im Kern geht es darum, die Denkfähigkeit eines Individuums zu testen (Turing, 1950). Drei Personen spielen das sog. „Imitation Game“:

– Ein Fragensteller („Interrogator“ oder „Judge“),
– eine Frau („Confederate“),
– ein Mann („Competitor“).

Der Interrogator kennt seine Mitspielenden nur als X bzw. Y und kann sie weder sehen, noch hören. Nachrichten werden über einen Teleprompter ausgetauscht. Durch Befragung soll der Interrogator herausfinden, wer von ihnen die Frau ist. Während die Frau dem Interrogator helfen möchte, versucht der Competitor, ihn zu täuschen. Ob dieser dabei Erfolg hat, hängt ganz von seiner Denkfähigkeit ab.

Als Erweiterung ersetzt Turing den menschlichen Competitor durch einen Computer. Auch die Frau wird zum sog. Confederate generalisiert. Wenn nun ein Computer die Rolle des Competitors genau so gut wie ein Mensch spielen kann, dann muss er Turing zufolge auf eine gewisse Art und Weise denkend, also intelligent, sein.

Nach (Hingston, 2009) könne man den Test auch so verstehen, dass es darum gehe, menschlich zu wirken und Intelligenz ein Weg sei, dies zu tun. Insofern ist der Turing-Test für Bots ein Test der Glaubwürdigkeit des Bots.

Diese Glaubwürdigkeit kann in zwei Kategorien unterteilt werden (Togelius et al., 2012):

– Charakter-Glaubwürdigkeit („Character believability“):
  Jemand glaubt, dass der Bot selbst real, eine echte lebende Person, ist.
– Spielenden-Glaubwürdigkeit („Player believability“):
  Jemand glaubt, dass ein echter Mensch als dieser Bot spielt und nicht, dass er von einem Computer gesteuert wird.

Während die Charakter-Glaubwürdigkeit wohl eher im Sinne von Turings originalem Test wäre, fokussieren sich Turing-Tests für Bots auf die Spielenden-Glaubwürdigkeit. Dahinter steht der beschriebene Konsens, dass glaubwürdige menschliche Bots die Spielerfahrung verbessern (Ortega et al., 2013; Soni & Hingston, 2008; Togelius et al., 2012).

Zu den ersten Bot-Turing-Tests zählt der „2K BotPrize“ (Hingston, 2009). Im Rahmen des Tests treten zwei Menschen und ein Bot in einer zehnminütigen Runde Unreal Tournament 2004 gegeneinander an. Die Menschen nehmen die Rollen Interrogator und Confederate ein, während der Bot gleichbleibend als Competitor auftritt. Nach Abschluss der Runde muss ein Interrogator entscheiden, welcher der Mitspielenden ein Mensch war.

Als Entscheidungsgrundlage dienen dem Interrogator nur die Interaktionen im Spiel. Im Gegensatz zum originalen Test versucht die Confederate nicht mehr zu helfen. Sie will lediglich das Spiel gewinnen. Wenn sich ein Bot also im Spiel so verhalten hat, dass ihn der Interrogator nicht von einem menschlichen Mitspielenden unterscheiden kann, so hat er den Test bestanden.

Auch wenn keiner der Bots die notwendige Anzahl an Interrogatoren überzeugen konnte, wurden erfolgreiche Strategien zum Design von glaubwürdigen Bots und zum Enttarnen von Bots herausgefiltert.

Eine solche Strategie war, leichte Fehler ins Gameplay der Bots zu integrieren, um ihre maschinelle Perfektion zu verbergen. Einige Bots haben auch pseudo-zufällige Verhaltensweisen verwendet, um unvorhersehbarer zu sein, da diese Eigenschaft vorwiegend Menschen zugeordnet wurde.

Manche Interrogatoren entwickelten spezifische Strategien für einige der Testszenarien. Ihre menschlichen Mitspielenden konnten sich an diese anpassen, während die Bots versagten. Anpassungsfähigkeit war somit ein entscheidendes Kriterium.

In den Folgejahren wurde der 2K BotPrize mehrfach ausgetragen und im Jahr 2010 noch einmal angepasst (Hingston, 2010). Wesentlicher Unterschied war die Einführung einer speziellen Waffe, die den Bewertungsprozess ins Spiel integrierte. Sobald ein Interrogator davon überzeugt war, einen Bot vor sich zu haben, konnte er diesen mit der besagten Waffe angreifen, was den Bot sofort erledigte. Sollte es sich allerdings nicht um einen Bot handeln, stirbt der Interrogator selbst.

Diese Änderung wird unter anderem von (Togelius et al., 2012) kritisiert. Mit der Einführung dieser Waffe werde klar, dass sich einige Interrogatoren während der Tests mehr auf das Spielen, als auf das Bewerten der Bots konzentrieren und umgekehrt. Abgesehen davon erfordere die mehr oder minder aktive Teilnahme am Spiel von jedem Interrogator ein Mindestmaß an Konzentration. Diese Konzentration ist begrenzt und sollte lieber auf die Bewertung angewendet werden. Ähnliches wurde bereits zuvor von (Laird & Duchi, 2000) angemerkt.

Togelius et al. empfehlen, eine Bewertung erst nach dem Spielen abzugeben. Zudem ist zu diskutieren, ob für eine Bewertung überhaupt die Teilnahme am Spiel erforderlich ist.

Entsprechend haben mehrere Bot-Turing-Tests den Interrogator aus dem Test-Szenario entfernt (Devlin et al., 2021; Laird & Duchi, 2000; Milani et al., 2023; Ortega et al., 2013). Anstatt selbst am Spiel teilzunehmen, beobachtet er lediglich Aufnahmen der Test-Szenarien. Dadurch kann er sich vollends auf die Bewertung konzentrieren.

Gleichzeitig wird dadurch auch die Perspektive gewechselt. Die Interrogatoren verfügen nun über den gleichen Informationsstand, wie der Competitor, da sie das Spiel aus seiner Sicht sehen. Dies ist vor allem für Spiele wie Skat vorteilhaft, in denen es keine perfekten Informationen gibt.

Es wird ebenfalls die Beobachtungszeit maximiert. (Laird & Duchi, 2000) weisen beispielsweise darauf hin, dass Bots in vielen kompetitiven Spielen nur innerhalb der kurzen Duelle beobachtet werden können, wofür sich die Spielteilnehmenden erst einmal begegnen müssen. Ein Test-Aufbau wie im 2K Bot bewirkt kurze Interaktionen mit den Bots. Wo möglich, sollte also die Perspektive des Testsubjekts gezeigt werden.

Abschließend haben sich folgende Modi für die Beurteilung ergeben (Togelius et al., 2012):
– Vergleichend (Milani et al., 2023; Ortega et al., 2013),
– Boolsch (Hingston, 2010; Laird & Duchi, 2000; Togelius et al., 2012).

In vergleichenden Tests werden die Videos von zu beurteilenden Bots mindestens paarweise gegenübergestellt. Im Vergleich dazu wird bei boolschen Tests nur eine Ja/Nein-Frage zu einzelnen Videos beantwortet. Bis jetzt gibt es keinen Konsens darüber, welche Methode sich mehr eignet. Beide Methoden haben Vor- und Nachteile, die in (Togelius et al., 2012) diskutiert werden.

Im nächsten Kapitel wird auf das Design des in dieser Arbeit durchgeführten Bot-Turing-Tests eingegangen.

3. Methodik

Als Rahmen für den durchgeführten Quasi-Turing-Test wurde ein Fragebogen mit Video gewählt. Dieser war online über Google Formulare verfügbar und wurde innerhalb von vier Tests mit jeweils anderem Video genutzt. Die Teilnehmenden stammten dabei hauptsächlich aus dem Skatfreunde-Discord-Server, wo der Test als viertägiges Event abgehalten wurde.

In den einzelnen Unterkapiteln wird das genaue Design der Studie, die Generierung von Videomaterial und die Durchführung des Tests beschrieben.

3.1. Studiendesign

Allem voran soll der Fragebogen bei der Beantwortung der Forschungsfragen helfen, indem er folgende Punkte behandelt:

- Wie qualifiziert erscheinen Teilnehmende für die Beurteilung,
- Was ist ihr Urteil bezüglich Spielstärke, Kooperation und Glaubwürdigkeit der Bots,
- Welche Beobachtungen haben sie für ihre Beurteilungen genutzt.

Deshalb stehen am Anfang des Bogens demografische Fragen, die Erfahrung und Fähigkeitsniveau der Testpersonen einordnen sollen. Diese lauten wie folgt:

- **Wie erfahren würdest du dich selbst einschätzen?**
  Eine 5-Punkte Likert-Skala von „Ich kenne die Regeln [von Skat]“ bis „Ich bin Profi“.
- **Wie viele Punkte holst du durchschnittlich pro Liste?**
  Freitext-Antwort; Quantitative Untermauerung der Selbsteinschätzung.
- **Bist du in einem Verein oder nimmst an Turnieren teil?**
  Optionale Multiple Choice, um sich ggf. noch weiter zu abzugrenzen; Hier soll dargestellt werden, wie ernst die Befragten Skat nehmen.
- **Wie oft spielst du Skat gegen den Computer?**
  Eine 5-Punkte Likert-Skala von „Ich spiele am meisten mit anderen Personen“ bis „Ich spiele am meisten gegen den Computer“.

Entsprechend der vorgestellten Kritik an Tests, wo Beurteilende selbst am Spiel teilnehmen, wurde sich für Videoaufnahmen entschieden. Diese folgen im zweiten Teil.

Die Videos zeigen eine Runde Skat ab Beginn des Stichspiels. Zu sehen ist die Perspektive des Testsubjekts. Die Reizwerte werden ergänzend zur Verfügung gestellt, da sie ggf. Aufschluss über das Blatt des AS geben, was wiederum für Aktionen der GS relevant ist.

Der Einfachheit halber wurde für diese Studie der boolsche Test-Modus gewählt. Eine vergleichende Studie erfordert größeren Umfang, da mit jedem evaluierten Bot die Anzahl an miteinander zu vergleichenden Bots steigt. Da es nicht abzusehen ist, wie groß die Beteiligung der genannten Discord Community ausfällt, wurde sich entschlossen den Umfang und damit auch den Aufwand für Teilnehmende zu minimieren. Dadurch soll eine möglichst hohe Beteiligung erreicht werden.

Nach dem Video und der eröffnenden Frage „Wurde die Partie von einem Menschen gespielt?“ ist das Urteil anschließend in einem Freitextfeld zu begründen. Aus den Antworten in diesem Feld sollen später Merkmale extrahiert werden, die zur Beantwortung von Forschungsfragen drei und vier genutzt werden.

Darauf folgen 5-Punkte Likert-Skalen um Spielniveau und Kooperation einzuschätzen:

- „Wie hoch schätzt du das Niveau des [betrachteten] Spielers ein?“,
- „Wie gut haben die Gegenspieler zusammengearbeitet?“.

Die Skalen helfen bei der Beantwortung von Forschungsfrage zwei.

Abschließend wird mit einer weiteren Freitextantwort erfragt, woran sich die Befragten bei der Bewertung der Kooperation orientiert haben. Es sollen Merkmale für gute bzw. schlechte Kooperation gefunden werden, welche Forschungsfrage vier beantworten.

3.2. Generierung von Videomaterial

Wie das Videomaterial im Zentrum des Tests entstanden ist, soll in diesem Kapitel erläutert werden. Es wurde sich für folgende Quellen von Testsubjekten entschieden:

– **„Skat Lernen“**
  YouTube-Kanal (D. Schäfer, 2016); Für Spiele mit menschlichen Experten.
– **„Skat“ von Isar Interactive GmbH & Co. KG** (Isar Interactive GmbH & Co Kg, 2014)
  Marktführer im Appstore.
– **„Skat am Stammtisch“ von StammtischGames GmbH & Co. KG** (StammtischGames GmbH & Co. KG, 2015)
  Top 5 für offline Skat-Spiele im Appstore.
– **„Skatfreunde“ von Bestjack Entertainment GmbH**
  Arbeitet mit „Fox“ KI von Edelkamp et. al. (Edelkamp, 2021a; Gößl, 2019)

Diese Quellen sind ausschließlich kommerzieller Natur. Akademische KIs waren während der Recherche schwer bis gar nicht zugänglich. Es gibt zwar einen Server von (Buro, 2007), auf dem man gegen XSkat und Versionen von Kermit spielen kann, allerdings erschweren Interface sowie Spielerfahrung qualitativ hochwertige Aufnahmen. Ein Zeitlimit für jede Partie kommt erschwerend hinzu. Deswegen wurden akademische Testanwendungen als Quellen verworfen.

Eine Schwierigkeit für den Test ist, dass keine Möglichkeit besteht, die Kartenverteilungen zu bestimmen. Entsprechend wurde versucht, andere Parameter anzugleichen:

– In jeder App wurde eine Liste aus 36 Spielen durchgespielt und aufgenommen,
– Vom YouTube-Kanal wurde ein Livestream mit einer solchen Liste ausgewählt,
– Spezialspiele (Grand, Null) wurden aussortiert,
– Gewinnstufe „Schwarz“ wurde aussortiert.

Es wurde sich auf einfache Farbspiele beschränkt, da Grand oder höhere Gewinnstufen von KIs ohnehin weniger häufig angesagt werden (Keller & Kupferschmid, 2008). Zusätzlich ist bereits bekannt, dass die meisten KIs im Nullspiel schwach sind (Edelkamp, 2021a). Die Gewinnstufe „Schwarz“ wurde aussortiert, da dort alle Stiche an den AS gehen. Meist besteht in solchen Runden eine chancenlose Verteilung für die GS, weswegen Schneider-Schwarz-Runden wenig ertragreich für den Test erscheinen.

Um die spielerische Qualität der Auswahl an Aufnahmen zu optimieren, wurde Daniel Schäfer, Betreiber des YouTube-Kanals „Skat Lernen“ und Mitgründer von Bestjack Entertainment zu Rate gezogen. Er wählte je eine Partie aus, die seiner Einschätzung nach die meisten interessanten Entscheidungen für das später zu sehende Testsubjekt bot.

Für die Aufnahmen wurden folgende Grupen verwendet:

– Drei Menschen,
– Eine KI zusammen mit einem Menschen gegen eine weitere KI,
– Drei KIs.

Eine vierte mögliche Konstellation von einem Team aus Mensch und KI gegen einen weiteren Menschen ist vermutlich äquivalent zu Konstellation Nummer zwei, da sich lediglich der AS ändert, welcher nicht Teil der Beurteilung ist. Über die Äquivalenz hinaus gab es technische Limitationen, welche die besagte Konstellation schwer durchführbar machen. Beispielsweise bietet ausschließlich die Quelle von Isar Interactive eine Möglichkeit, mit zwei Menschen und einem Bot zu spielen. Aus diesen Gründen wurde stattdessen zwei Mal in der Konstellation Mensch plus zwei Bots aufgenommen.

Die Konstellation aus drei Menschen wird durch eine Live-Partie von Experten abgedeckt, die vom Skat-Lernen YouTube-Kanal stammt. Zur Auswahl der Partie wurden die oben beschriebenen Kriterien angewendet.

Allgemein wurden variable Konstellationen gewählt, um mehr Merkmale für das Gegenspiel gewinnen zu können. Das Ganze basiert auf dem Wissen, dass für Menschen Regeln beim Zusammenspiel existieren, welche den Bots nicht unbedingt geläufig sind.

Schlussendlich wurde von jeder ausgewählten Aufzeichnung der Spielverlauf in Skatfreunde nachgestellt und daraus ein Video für den Test erstellt. Jedes Video geht von der Ansage des Spiels durch den AS bis hin zur Auswertung, bei der die gesammelten Punkte gezählt und Sieg bzw. Niederlage verkündet werden.

Wie bereits angedeutet wurde das Reizen für die Aufnahmen übersprungen, um bei kurzer Videolänge ausreichend viel Zeit für das Stichspiel zu haben.

Generell wurde Interaktions- bzw. Beobachtungszeit von (Togelius et al., 2012) weitläufig diskutiert. Schlussendlich befinden sie, dass für eine Bewertung die Zeit eines Levels bzw. einer Runde genügt. Dementsprechend sind alle im Test gezeigten einzelnen Runden gleich lang.

Die einheitliche Videolänge von 45 Sekunden resultiert hauptsächlich aus der genormten Ausspielzeit für jede Karte. Geschwindigkeit war in den meisten vorangegangenen Tests ein entscheidener Faktor für die Unterscheidung zwischen Mensch und Maschine (Ashktorab et al., 2020; Milani et al., 2023; Ortega et al., 2013). Daher wurde eine einheitliche Ausspielgeschwindigkeit gewählt, um lange Bedenkzeiten der Bots zu verbergen.

Spielstärke von Bots korreliert oftmals stark mit Rechenzeit. Entsprechend wurde die leicht schnellere, einheitliche Spieldarstellung gewählt, um den bereits bekannten Unterscheidungsfaktor Zeit zu maskieren.

Abschließend wurde Skatfreunde als Plattform für die Videos gewählt, um den Störfaktor Präsentation zu entfernen. In den Videos ist also immer die gleiche Oberfläche zu sehen. Es ist ebenfalls davon auszugehen, dass Skatfreunde jeder teilnehmenden Person bekannt ist, da die Testenden vom Skatfreunde-Community-Discord stammen. So können Missverständnisse bezüglich des Spielgeschehens ausgeschlossen werden.

Was genau der besagte Discord ist und wie der Test dort abgehalten wurde, ist im nächsten Unterkapitel beschrieben.

3.3. Durchführung

Für die Rekrutierung von Testenden wurde der Community Discord von Skatfreunde genutzt. Ein „Discord“ ist ein Server in der gleichnamigen App (Discord Inc., 2024). Auf solchen Servern gibt es unterschiedliche Kanäle für Text- oder Sprach- und Videochat. Jeder Discord-Nutzende kann einen solchen Server selbst erstellen und andere dorthin einladen.

Der Server von Skatfreunde dient hauptsächlich dazu, die Entwicklung der namensgebenden App „Skatfreunde“ zu begleiten. Auf dem Server werden Updates zum Spiel und Nachrichten an die Gemeinschaft geteilt. Darüber hinaus existieren auf dem Server auch diverse Themenkanäle, wo Skat diskutiert wird, oder Sprachkanäle, wo Mitglieder gemeinsam Skat spielen.

Entsprechend findet sich auf dem Server eine breite Masse von Skat-Interessierten, die vom Hobbyspielenden bis zum aktiven Vereinsmitglied reicht. Deswegen wurde er auch als Plattform zur Veröffentlichung der Tests gewählt. Neben anwesenden spielstarken Experten werden somit gesichert Personen erreicht, die mindestens die Grundregeln des Spiels beherrschen.

Der gesamte Test wurde über einen Zeitraum von vier Tagen durchgeführt, wobei an jedem Tag ein Fragebogen ausgefüllt wurde. Die Verteilung der Fragebögen erfolgte im Sinne der Server-Besitzenden, um die Anzahl an Benachrichtigungen pro Tag möglichst im Rahmen zu halten. Zudem ist so für eine kontinuierliche Bereitstellung neuer Inhalte auf dem Server gesorgt.

Die Reihenfolge der Test-Konstellationen wurde zufällig gewählt und sah wie folgt aus:
- Tag 1: Spiel von menschlichen Experten; Drei Menschen,
- Tag 2: Experte spielt in Skat am Stammtisch; Bot mit Mensch gegen Bot,
- Tag 3: Experte spielt in Skat von Isar Interactive; Bot mit Mensch gegen Bot,
- Tag 4: Spiel der Skatfreunde-KI gegen sich selbst; Drei Bots.

Der erste Tag begann mit einer Ankündigung der Testreihe und anschließender Veröffentlichung des ersten Tests während der Mittagszeit. Jeder weitere Test wurde zu dieser Tageszeit veröffentlicht.

Zu Beginn eines Tests erfolgt ein Briefing, welches kurz den Test beschreibt, die Rahmenbedingungen des Tests erklärt und klare Aufgaben an die Testenden stellt. Im Test werde davon ausgegangen, dass:
- Der betrachtete Spielende entweder Mensch oder KI sein kann,
- Die Identität der Mitspielenden unbekannt ist,
- Mitspielende immer ihr Bestes geben.

Anweisungen an die Testenden sind folgende:
- Achte auf die Spielweise des Testsubjekts,
- Entscheide, ob es sich um einen Menschen oder eine KI handelt,
- Bewerte nicht das Können der anderen Mitspielenden,
- Sondern bewerte, wie gut das Team miteinander gespielt hat.

Des Weiteren wird darauf hingewiesen, dass man sich das Video beliebig oft ansehen könne und kein Zeitdruck bestehe.

Am Ende des Fragebogens befindet sich eine Danksagung und die Bitte, den Test erst später zu diskutieren. Alle Mitglieder des Servers sind dieser Bitte stets nachgekommen, sodass keine Testperson offensichtlich von den Meinungen der anderen Testenden beeinflusst wurde. Die Diskussion wurde erst am Abend des Testtages durch einen weiteren Post freigegeben, auf den der Fragebogen dann geschlossen wurde.

Auf diese Art und Weise wurde über einen Zeitraum von vier Tagen täglich ein Test veröffentlicht, von den Servermitgliedern beantwortet, geschlossen und anschließend diskutiert. Erst nach Schließung des finalen Fragebogens wurden die Ergebnisse aller Tests auf dem Server bekannt gegeben.

4. Auswertung

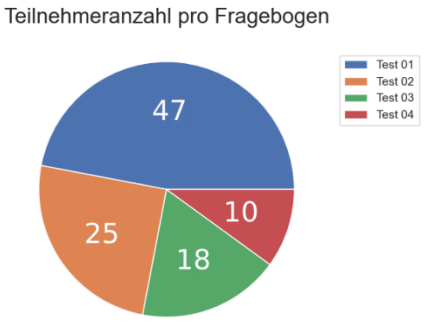


Abbildung 1: Teilnehmendenzahl des Skat-Turing-Tests.

Insgesamt wurden über die vier Fragebögen verteilt 100 Skatspielende befragt. Die genaue Aufteilung ist in Abbildung 1 zu sehen. Wie eingangs befürchtet, nahm sie mit jedem fortfolgenden Test ab.

Im Folgenden werden pro Test zuerst die erhobenen Daten zum Testsubjekt behandelt und danach die demografischen Daten der Testteilnehmenden. So soll zunächst die Frage nach der Glaubwürdigkeit beantwortet werden, während im gleichen Schritt wahrgenommenes Fähigkeitsniveau und Kooperation dargestellt werden, um sie im Kontext zu betrachten.

In den darauffolgenden Schritten soll mit Hilfe von qualitativen Merkmalen und demografischen Daten analysiert werden, wie Testurteile und Wahrnehmung zustande gekommen sind. Dabei wird insbesondere untersucht, welche Merkmale bei der Erkennung am häufigsten verwendet wurden. Welche Eigenschaften der Testteilnehmenden zur Erkennung wichtig sind, wird ebenfalls betrachtet.

Die demografischen Testdaten werden zur Analyse nach erfolgreicher Erkennung gruppiert. Wenn in einem Test tatsächlich ein Mensch zu sehen war, werden alle Fragebögen mit dem Testurteil „Ja“ der Gruppe „Korrekt“ zugeordnet. Daten aus dieser Gruppe haben in den folgenden Abbildungen immer die Farbe Blau. Alle Bögen mit „Nein“ werden entsprechend der Gruppe „Falsch“ zugeordnet, welche die Farbe Orange hat. Umgekehrte Zuordnung gilt für Tests mit einem Bot, wo „Nein“ die korrekte Antwort ist.

Jedes Testkapitel startet mit einem Kreisdiagramm, auf dem die Anzahl an Antworten zum entsprechenden Fragebogen (kurz „N“) zu sehen ist. Darauffolgende Diagramme gehen alle von dieser Stichprobengröße aus, sofern keine abweichende Angabe in der Bildunterschrift gemacht wird.

Bevor im Einzelnen auf die Bögen eingegangen wird, beschreibt der nächste Abschnitt, welche Merkmale aus der Befragung gewonnen wurden und wie diese gebildet wurden.

Merkmale

In diesem Unterabschnitt wird die Extraktion von Merkmalen aus den Freitextfeldern beschrieben. Befragte sollten in diesen Feldern ihr Urteil begründen. Dazu wurden zunächst alle Antworten gelesen und mit einem oder mehreren Schlagwörtern versehen. Diese Schlagwörter wurden dann so präzise wie möglich zu den finalen Merkmalen zusammengefasst. Final hat dann jede Antwort eines oder mehrere Merkmale erhalten.

Für die Begründung des Testurteils, also ob ein Mensch oder ein Bot gespielt hat, sind so sechs verschiedene Merkmale entstanden:

- **Intuition,**
  Die Person hat keinen erkennbaren Anhaltspunkt und/oder einfach aus Gefühl heraus geantwortet.
- **Konventionelle Spielweise,**
  Befragte sprachen von „Standardzügen", „Normales Spiel" und davon, dass sich an geläufige Skat-Merksätze bzw. Konventionen gehalten wird.
- **Unorthodoxe Spielweise,**
  Es wurde sich nicht an bekannte Konventionen gehalten oder unerwartet gehandelt, was nicht allein für gutes oder schlechtes Spiel steht.
- **Gute Züge,**
  Befragte loben aus unterschiedlichen Gründen einzelne Züge oder heben Aktionen positiv hervor.
- **Schlechte Züge,**
  Befragte kritisieren aus unterschiedlichen Gründen einzelne Züge oder weisen auf Fehler hin.
- **Geschwindigkeit.**
  Es wird davon gesprochen, dass „gleichmäßig" gespielt wurde oder, dass „Ausspiele schnell kamen"

Befragte sollten ebenfalls beschreiben, warum die GS gut bzw. schlecht gespielt haben. In ihren Antworten finden sich folgende Merkmale:

- **Volle,**
  Im Skat sind Ass und Zehn die sog. „Vollen". Befragte bewerteten, wie diese innerhalb der Partie eingesetzt wurden oder kritisierten deren Verlust.
- **Klären,**
  Vom Klären wird gesprochen, wenn eine Farbe beim AS als fehlend identifiziert werden konnte. Diese Information erlaubt, die Farbe zu vermeiden, um Trümpfe zu verhindern oder eben die Bedienregel auszunutzen, um Stiche zu gewinnen.
- **Nachspielen,**
  Nachspielen ist es, wenn man die gleiche Farbe ausspielt, die zuvor vom Mitspielenden angespielt wurde. Das ist eine grundlegende Strategie fürs Gegenspiel und soll bewirken, dass man dem unbekannten Plan des Teammitglieds folgt. Entsprechend ist nicht nachzuspielen eine Form von Misstrauen.
- **Stellungsspiel,**
  Befragte hoben hervor, wie mit den Positionen am Tisch umgegangen wurde. Innerhalb einer Partie ist es entscheidend, wann wer die erste Karte legt bzw. legen muss. Eine gängige Strategie ist es, den AS in Mittelhand zu bringen, um auf seine Karten reagieren zu können. Entsprechend wurde diese Vorgehensweise als gute Kooperation gelobt, während unpassende Positionswechsel gerügt wurden.
- **Schmieren,**
  Wertvolle Karten bei einem gewonnenen Stich beizugeben wurde gelobt. Die gegensätzliche Aktion „Schnippeln", also nicht die wertvollste Karte zu spielen, wurde protestiert.
- **Ausspiel,**
  Befragte nehmen Bezug auf die Karte, mit der ein Stich eröffnet wird. Sie kann zu hoch oder zu tief sein oder ihrer Meinung nach die richtige oder falsche Farbe anspielen.
- **Trumpf,**
  Die Handhabung von den Trumpfkarten der GS wurde gelobt bzw. kritisiert, insbesondere wenn Trumpf von ihnen ausgespielt wurde.

– **Spielausgang,**
  Manche Befragte bezogen sich in ihrer Bewertung explizit auf den Ausgang des Spiels, also in etwa wie „die Kooperation war gut, weil gewonnen wurde“ oder „es wurde schlecht zusammengespielt, weil verloren wurde“

Mit den sechs Merkmalen für Testurteil und acht für Kooperation kann nun die Auswertung des ersten Fragebogens beginnen.

4.1. Test 01 – Expertenrunde

Der erste Fragebogen beinhaltet eine Runde aus menschlichen Experten. Es wurde Kreuz mit einem Reizgebot von 18 gespielt. AS ist Vorhand, links vom gezeigten Spielenden in Hinterhand. Das Spiel wurde von dem AS mit 62 Augen knapp gewonnen.

Glaubwürdigkeit

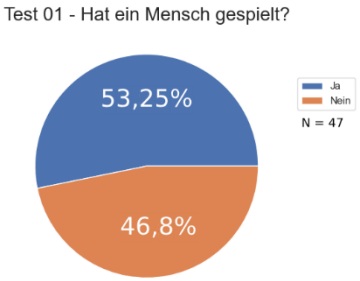


Abbildung 2: Testurteile in Test 01, korrektes Urteil „Ja“ in blau.

In Abbildung 2 sieht man, dass sich die 47 Befragten uneinig über die Identität des Testsubjekts sind. 25 Personen gaben korrekterweise an, einen Menschen beobachtet zu haben, was durch das blaue Stück des Diagramms dargestellt wird. Die 22 anderen Antworten in Orange hingegen glaubten, dass es sich um einen Bot handle. Im Sinne der Glaubwürdigkeit konnte der Experte die Hälfte der Befragten überzeugen.

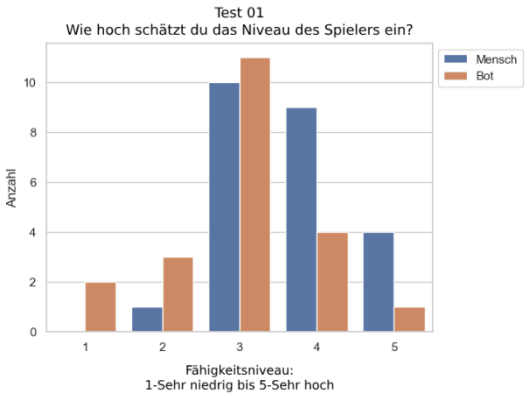


Abbildung 3: Wahrgenommenes Fähigkeitsniveau in Test 01.

Unter der Annahme, dass die Partie von einem Menschen gespielt wurde, wiesen Befragte dem Fähigkeitsniveau im Median vier von fünf Punkten zu. Man kann in Abbildung 3 ebenfalls sehen, dass der vermeintliche Mensch weniger niedrige Bewertungen für das Spiel erhielt als der vermeintliche Bot. Dieser kam im Median auf drei Punkte.

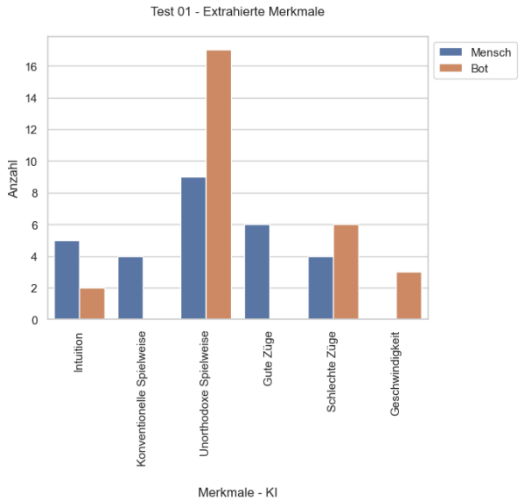


Abbildung 4: Merkmale aus Testurteilen von Test 01, mehrfaches Vorkommen möglich.

Die geringere Bewertung des Bots zeigt sich auch an den aufgetretenen Merkmalen. Abbildung 4 zeigt das Merkmal „Gute Züge“ ausschließlich bei Mensch-Antworten und nie bei Bot-Antworten. Erstere weisen darüber hinaus weniger „Schlechte Züge“-Merkmale auf.

In beiden Gruppen wurde unorthodoxe Spielweise am häufigsten zur Beurteilung herangezogen. Allerdings hat dieses Merkmal Befragte hauptsächlich in die Irre geführt. Am meisten tritt es in der falsch urteilenden Gruppe auf.

Abbildung 5: Wahrgenommene Kooperation in Test 01.

Abbildung 5 zeigt eine große Menge an Antworten mit vier von fünf Punkten für die Zusammenarbeit des Menschen. Der Bot erhält die meisten Antworten bei zwei Punkten. Im Median unterscheiden sich beide weniger stark, vier Punkte für die Kooperation des Menschen und drei Punkte für den Bot.

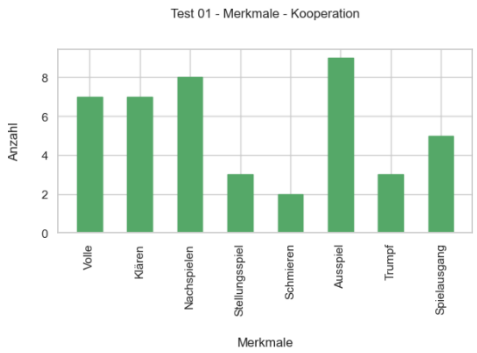


Abbildung 6: Merkmale für Kooperation in Test 01, N = 24, mehrfaches Vorkommen möglich.

In Abbildung 6 haben insgesamt die Hälfte der Befragten Merkmale zur Kooperation geliefert. Das „Ausspiel“ stellte für die Befragten den entscheidenden Faktor dar, weil in der gezeigten Partie viele Stiche untypischerweise nicht mit Assen begonnen wurden. Entsprechend ist das Merkmal „Volle“ ebenfalls häufig vertreten.

Als zweithäufigste Kritik wurde geäußert, dass das Testsubjekt an einer Stelle nicht Pik nachgespielt hat. Befragten zufolge könne man so noch die Farbe klären, was dieses Merkmal häufig auftreten lässt.

Zusammenfassend ist die Identität des Spielenden aus Test 01 unter den Testteilnehmenden umstritten. Die Glaubwürdigkeit wird mit 53% bewertet. Bei den bewerteten Kategorien „Fähigkeitsniveau“ und „Kooperation“ wurden höhere Bewertungen vergeben, wenn die Befragten der Meinung waren, dass es sich beim Testsubjekt um einen Menschen handelt.

In den Merkmalen zum Testurteil zeigt sich eine ähnliche Tendenz. Der Mensch wird in diesem Test mit mehr positiven Merkmalen assoziiert. Allerdings ist das häufigste Merkmal auch das irreführendste, da die unorthodoxe Spielweise die meisten Befragten dazu gebracht hat, den Spielenden für einen Bot zu halten.

Diese unorthodoxe Spielweise wurde auch in den kooperativen Merkmalen beschrieben, wo die häufigsten Merkmale „Ausspiel“, „Nachspielen“ und „Klären“ in negativer Weise aufgetreten sind. Sie wurden von Befragten kritisiert oder als komisch bezeichnet.

Als nächstes werden die demografischen Eigenschaften der Testteilnehmenden betrachtet.

Demografie

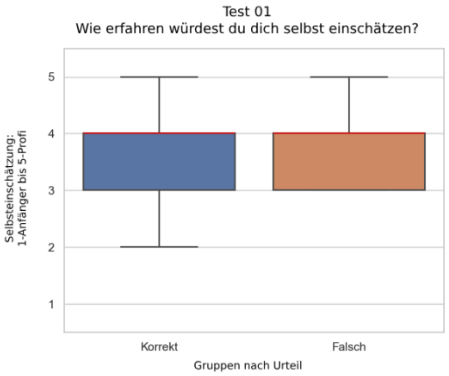


Abbildung 7: Selbsteinschätzung der Testpersonen aus Test 01.

In Abbildung 7 ist ersichtlich, dass Testteilnehmende der Gruppen sich im Median für gleich gut halten. Beide liegen bei einer vier auf der Skala. Das Minimum ist bei der korrekt urteilenden Gruppe einen Punkt niedriger als bei der falsch urteilenden. Abgesehen davon ist die Verteilung identisch.

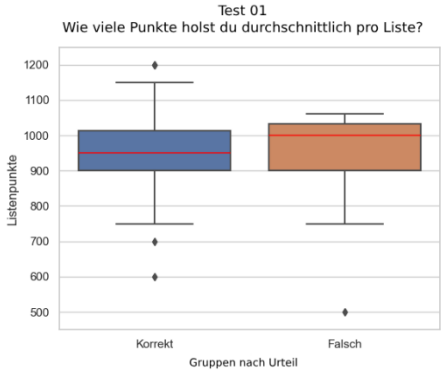


Abbildung 8: Listenpunkte aus Test 01, N = 39.

Die Angabe der Listenpunkte ist in Abbildung 8 zu sehen. Da es sich bei dem Eingabefeld im Fragebogen um ein Freitext-Feld handelte, kam es auch zu ungültigen Antworten. Beispielsweise wurden unrealistische Werte eingetragen, weil von einer anderen Listengröße ausgegangen wurde. In anderen Fällen wurde Text oder schlichtweg gar nichts eingegeben.

Für Abbildung 8 und alle fortfolgenden Auswertungen der Listenpunkte, wurden deshalb ungültige Antworten herausgefiltert. Somit haben in Test 01 insgesamt 39 Personen ihre durchschnittlichen Listenpunkte mitgeteilt.

Es lässt sich feststellen, dass die Gruppe „Falsch“ im Mittel eine höhere Anzahl an Listenpunkten aufweist als die andere Gruppe. Während beide Gruppen das gleiche Minimum haben, ist das Maximum der Gruppe „Korrekt“ höher. Beide Gruppen verfügen über Ausreißer: 1200, 700, 600 für die korrekt urteilende Gruppe und 500 für die falsch urteilende Gruppe. Abseits davon überlappen sich die Boxen weitestgehend.

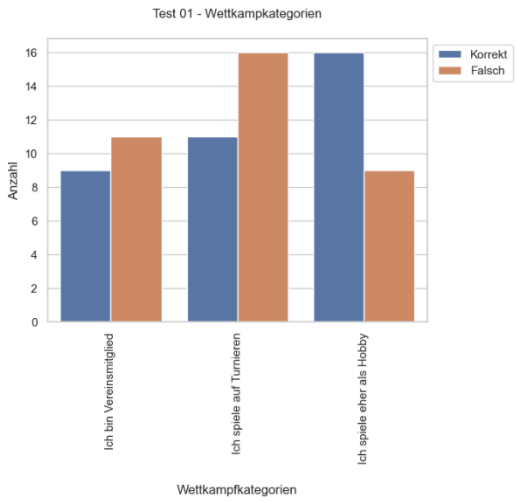


Abbildung 9: Wettkampfkategorien in Test 01, Mehrfachnennung möglich.

Die Einordnung in eine oder mehrere Wettkampfkategorien kam nach den Listenpunkten. Wie man Abbildung 9 sehen kann, besteht die Gruppe „Korrekt“ hauptsächlich aus Hobbyspielenden. Die Gruppe „Falsch“ tendiert dazu, auf Turnieren zu spielen.

Allgemeiner ist festzustellen, dass die Gruppe „Falsch“ tendenziell kompetitiver ist. Sie enthält mehr Antworten der Kategorien Turnier und Verein als die andere Gruppe.

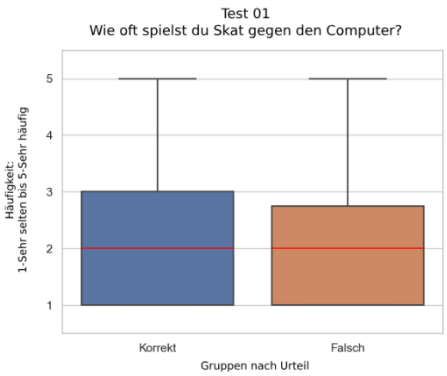


Abbildung 10: Erfahrung mit Bots aus Test 01.

Abbildung 10 veranschaulicht die Erfahrungen der Gruppen im Spielen mit Bots. Die Verteilungen fallen nahezu identisch aus. Beide Gruppen haben im Median eher wenig Erfahrung (Wert „2“). Es wird in der Regel mit anderen Menschen zusammen gespielt.

Abschließend lässt sich also aus den demografischen Daten lesen, dass die Gruppe „Falsch“ im Median über bessere Spielenden verfügt und auch kompetitiver ist. Währenddessen weist sie die gleiche Menge an Erfahrung wie die andere Gruppe auf.

4.2. Test 02 – Skat am Stammtisch

Im zweiten Test war eine Runde aus Skat am Stammtisch zu sehen. Es wurde die Perspektive eines Bots mit menschlichem Teammitglied gezeigt. Vorhand ist erneut AS mit 18 und spielt Karo. Der Bot sitzt in Mittelhand. Das Spiel wird vom AS mit 93 Augen und der Gewinnstufe Schneider gewonnen.

Glaubwürdigkeit

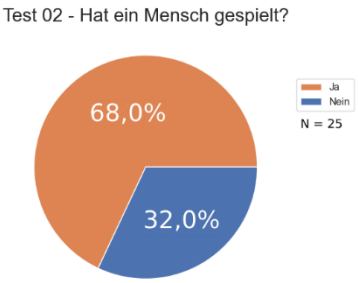


Abbildung 11: Testurteile in Test 02, korrektes Urteil „Nein“ in blau.

Aufgrund der gezeigten Partie sind sich 17 von 25 Befragten sicher, dass es sich um einen menschlichen Spielenden gehandelt haben muss. Wie man in Abbildung 11 sieht, liegen sie damit falsch. Lediglich acht Personen haben richtig geurteilt. Für die Glaubwürdigkeit des Bots heißt das, er konnte zwei Drittel der Leute täuschen.

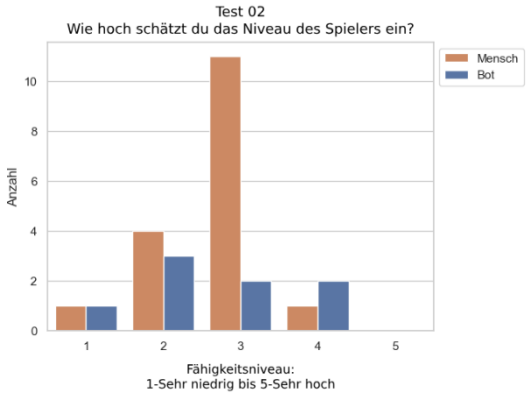


Abbildung 12: Wahrgenommenes Fähigkeitsniveau in Test 02.

Abbildung 12 zeigt, dass der Bot jedoch weniger gut bewertet wurde als der Mensch. Der Median für den Bot liegt bei zweieinhalb Punkten, während der Mensch eine Bewertung von drei erhält.

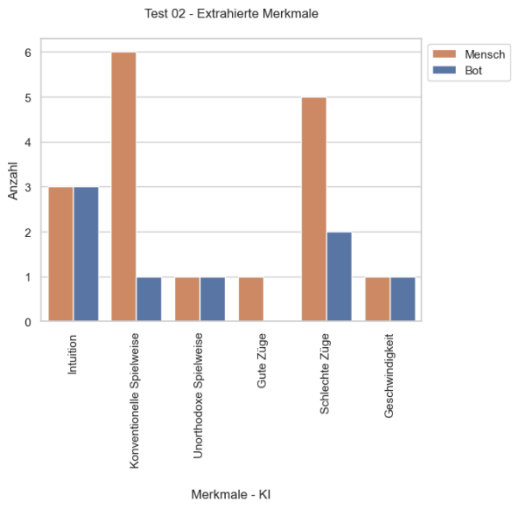


Abbildung 13: Merkmale aus Testurteilen von Test 02, mehrfaches Vorkommen möglich.

Wie in Abbildung 13 erkennbar ist, wurden hauptsächlich konventionelle Spielweise und schlechte Züge zur Identifikation eines Menschen genutzt. Schlechte Züge wurden ebenfalls häufig verwendet, um für einen Bot zu argumentieren. Allerdings erhält der vermeintliche Mensch im Verhältnis mehr Urteile mit schlechten Zügen als der Bot.

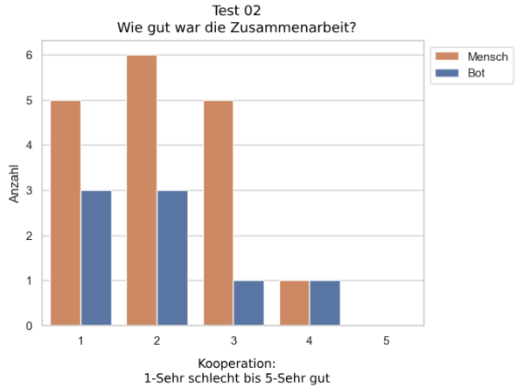


Abbildung 14: Wahrgenommene Kooperation in Test 02.

Neben den schlechten Zügen wird zusätzlich Menschen und Bots schlechte Kooperation bescheinigt. Abbildung 14 zeigt eine ähnliche Verteilung der Bewertungen. Beide Gruppen haben im Median zwei von fünf Punkten für die Zusammenarbeit erhalten.

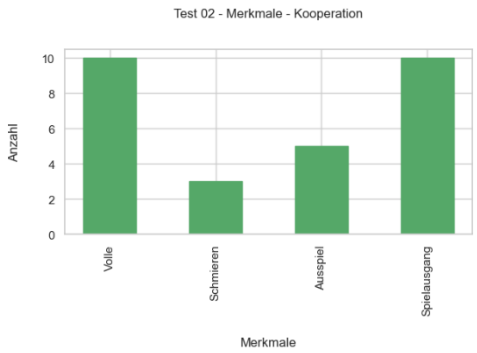


Abbildung 15: Merkmale für Kooperation in Test 02, N = 16.

16 von 25 Befragten lieferten Merkmale zur Kooperation. Abbildung 15 zeigt, dass bei zwei Drittel der „Spielausgang“ Einfluss auf ihre Bewertung genommen hat. Man hätte das Spiel nicht Schneider verlieren müssen, weswegen Befragte niedrig urteilten.

Die Ursache für diese unnötig harte Niederlage wird in der schlechten Kooperation gesehen. Der spielende Bot wird aber nicht von allen dafür verantwortlich gemacht. Sein (menschlicher) Mitspieler hielt seine Vollen zurück und verlor diese später, was einen Testteilnehmenden besonders verärgerte:

„Schneider verkaspert, den hätte man durch Pik Ass statt K sichern können. Das Spiel stand ohnehin sehr schlecht und ein Sieg war nicht mehr wirklich möglich. […] Auf den eigenen Mann schnippeln, unglaublich. Das gäbe am Tisch bestimmt Ärger.“

Andererseits ist dieser Fehler vielleicht auch durch das fehlende Ausspielen der eigenen Vollen in Herz begünstigt worden. So haben jedenfalls andere Teilnehmende geurteilt:

„[…] Gewinn ist nicht mehr möglich. Es geht nur um Schneiderfrei. Um einen Fehler des Mitspielers zu vermeiden, muss hier der eigene Volle auf den Tisch.“

„Im sechsten Stich lässt der Spieler seinem Partner die Chance dem Alleinspieler den Schneider zu schenken, wie es dann auch tatsächlich geschah.“

Entsprechend entscheidend war das Merkmal „Volle“ für die Beurteilung der Kooperation in diesem Test.

Insgesamt wurde in Test 02 eine Zweidrittelmehrheit für die Identität Mensch erzielt, nur war dies das falsche Urteil. Der Bot von Skat am Stammtisch hat folglich in der gezeigten Partie sehr glaubwürdig gespielt. Auch wenn seine Züge als schlecht bis konventionell wahrgenommen wurden, ging die Mehrheit der Personen davon aus, dass eine KI besser gespielt hätte.

Im Folgenden wird wieder ein Blick auf die Eigenschaften der Urteilenden geworfen.

Demografie

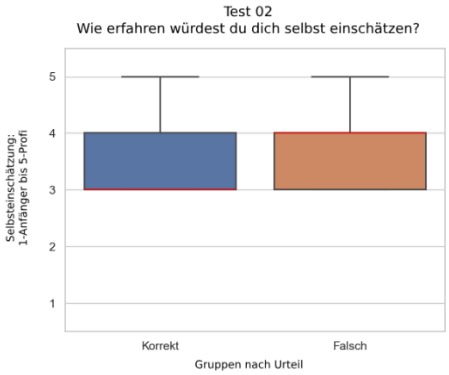


Abbildung 16: Selbsteinschätzung der Testpersonen aus Test 02.

Abbildung 16 zeigt gleiche Verteilung der Selbsteinschätzungen. Die Gruppe „Korrekt“ hat sich dabei im Median einen Punkt schwächer geschätzt als die andere.

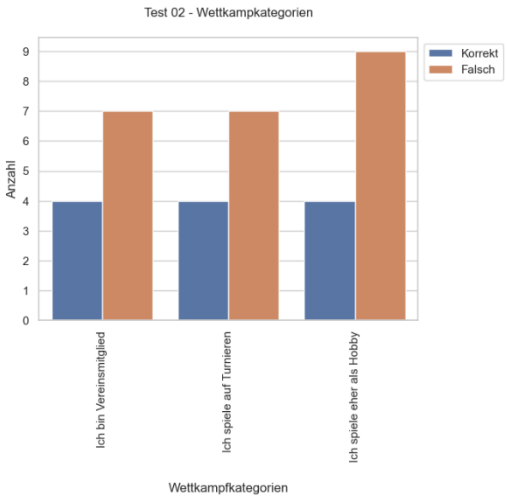


Abbildung 17: Wettkampfkategorien aus Test 02, N = 35.

Die Wettkampfkategorien sind in diesem Test fast gleich verteilt. Wie in Abbildung 17 zu sehen ist, haben im Verhältnis mehr Hobbyspielende falsch geurteilt.

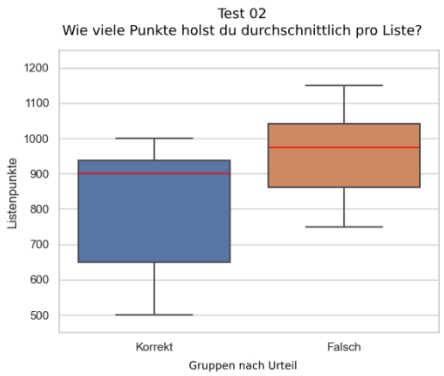


Abbildung 18: Listenpunkte aus Test 02, N = 23.

In Test 02 haben 23 Befragte ihre durchschnittlichen Listenpunkte mitgeteilt. Abbildung 18 weist einen ähnlichen Abstand der Mediane wie in Test 01 auf.

Die Boxen überlappen sich in diesem Test weniger stark. Es ist zu sehen, dass die Gruppe „Korrekt“ deutlich mehr in den niedrigen Punktebereich streut. Ihr Punktebereich ist tatsächlich so niedrig, dass die Person mit 500 Punkten kein Ausreißer ist.

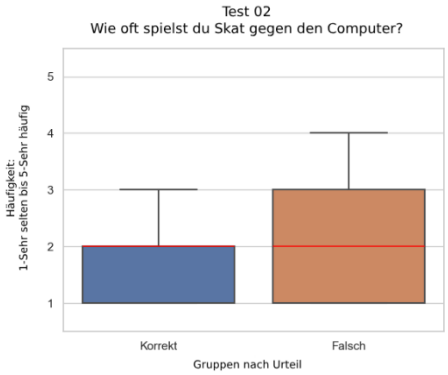


Abbildung 19: Erfahrung mit Bots aus Test 02, N = 25.

Erfahrung im Spielen mit Bots kann man in Abbildung 19 sehen. Beide Gruppen sind im Median gleich erfahren. Gruppe „Falsch“ weist dieses Mal einige erfahrene Spielende auf, was man an der höheren Box und dem höheren Maximum des Diagramms ablesen kann.

Das Fazit zum demografischen Teil von Test 02 fällt ähnlich aus wie im ersten: Die falsch urteilende Gruppe hat die besseren Skatspielenden und beide Gruppen sind im Median ähnlich unerfahren im Spielen mit Bots.

4.3. Test 03 – Skat von Isar Interactive

Test 03 zeigt erneut einen Bot mit menschlichem Teammitglied, dieses Mal in der App von Isar Interactive. Hinterhand hat mit 18 das Reizen gewonnen und spielt Karo. Der Bot sitzt in Vorhand. Der AS verliert mit 57 Augen knapp das Spiel.

Glaubwürdigkeit

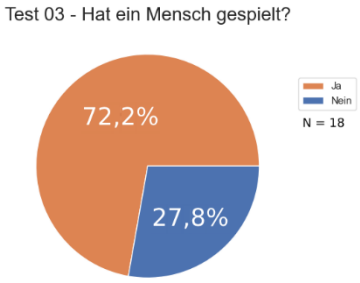


Abbildung 20: Testurteile in Test 03, richtiges Urteil „Nein“ in blau.

Wie man in Abbildung 20 sehen kann, fällt das Urteil noch ein wenig eindeutiger aus als im vorangegangenen Test. Ganze 13 der 18 Befragten wurden durch den Isar-Bot getäuscht. Lediglich fünf haben ihn korrekterweise als Bot erkannt. Insofern ist er noch glaubwürdiger als der Bot von Skat am Stammtisch.

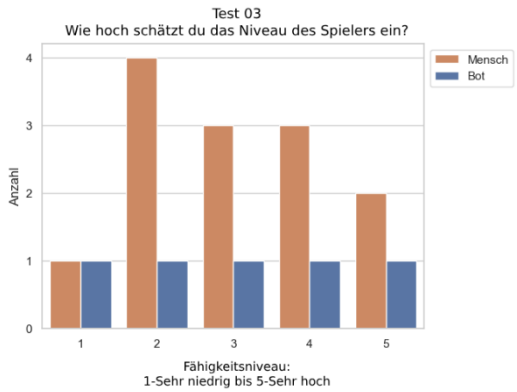


Abbildung 21: Wahrgenommenes Fähigkeitsniveau in Test 03.

In Abbildung 21 bekommt der vermeintliche Mensch von den meisten Befragten eine zwei von fünf für sein Spiel. Im Median sind beide Seiten mit drei von fünf bewertet worden.

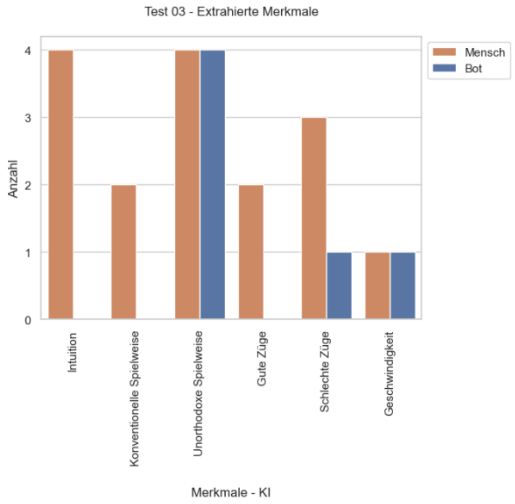


Abbildung 22: Merkmale aus Testurteilen von Test 03, mehrfaches Vorkommen möglich.

Wie man in Abbildung 22 sieht, wurde am häufigsten „unorthodoxe Spielweise“ als Begründung für die Urteile verwendet. Entscheidend war das Spielen einer Zehn an Stelle eines Asses, was von beiden Gruppen für ihr Urteil verwendet wurde. Dem Menschen wurden darüber hinaus gute und schlechte Züge zugeschrieben, während der Bot nur schlechte Züge als Merkmal aufweist.

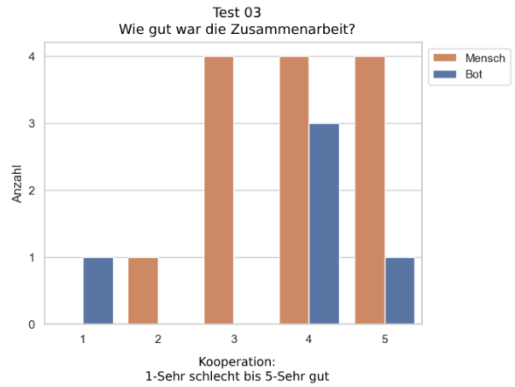


Abbildung 23: Wahrgenommene Kooperation in Test 03.

Abbildung 23 vergibt hohe Wertungen für Kooperation in diesem Test. Beide Gruppen landen bei vier von fünf.

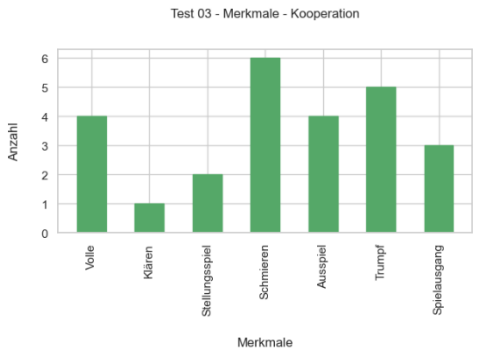


Abbildung 24: Merkmale für Kooperation in Test 03, N = 12, mehrfaches Vorkommen möglich.

12 von 18 Befragten gaben eine Erklärung für ihre Bewertung der Kooperation ab. Abbildung 24 zeigt, dass am meisten über „Schmierung“ gesprochen wurde. Der bereits zuvor erwähnte, unorthodoxe, Zug war eben genau so eine Situation. Die meisten Befragten hätten sich hier ein Ass als Schmierung gewünscht, weil Pik bereits geklärt worden war und man mit der gespielten Kreuz Zehn stattdessen noch einen Stich machen könne. Im Grunde genommen ging es bei der Schmierung also ebenfalls um die Verwendung der Vollen.

Das Merkmal „Trumpf“ tritt in diesem Spiel auf, weil der menschliche Mitspielende einen sog. „Trumpfabzug“ vorgenommen hat. Er eröffnete selbst mit Trumpf, um die Menge an Trümpfen des AS zu verringen.

Einer Hand voll Befragten hat es schon gereicht, dass die GS das Spiel gewonnen haben, um ihnen eine vier oder fünf für Kooperation zu geben. Ein Sieg der GS ist im Grunde genommen ein seltener Fall und in der gezeigten Partie nur möglich, weil das Blatt ohnehin schlecht für den AS stand. Alle verbliebenen Trümpfe, bis auf einen, waren beim gleichen Gegenspielenden.

Abschließend wurde in Test 03 wieder mehrheitlich falsch geurteilt. Die meisten Testteilnehmenden wurden durch den Bot getäuscht. Anteilig ist dieser Bot sogar erfolgreicher darin als der Vorherige. Aufgrund der technisch schlechten Schmierung wurde eher ein Mensch vermutet als eine Maschine.

Es folgen nun die demografischen Eigenschaften der Testteilnehmenden.

Demografie

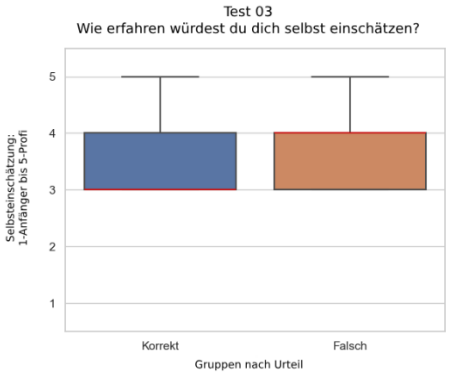


Abbildung 25: Selbsteinschätzung der Testpersonen aus Test 03.

Abbildung 25 hat die gleiche Verteilung für Selbsteinschätzung wie in Test 02. Die Medi-ane der Gruppen unterscheiden sich um einen Punkt.

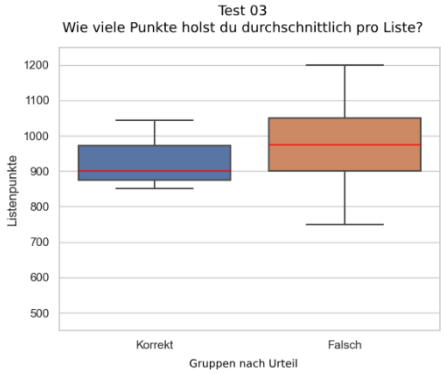


Abbildung 26: Listenpunkte aus Test 03, N = 16.

16 Personen gaben ihre Listenpunkte an, dargestellt in Abbildung 26. Die Gruppe „Falsch“ liegt erneut in Punkten vorne, denn ihr Median ist höher. Dieses Mal ist sie breiter gestreut und enthält sowohl den höchsten als auch den niedrigsten Wert des Tests.

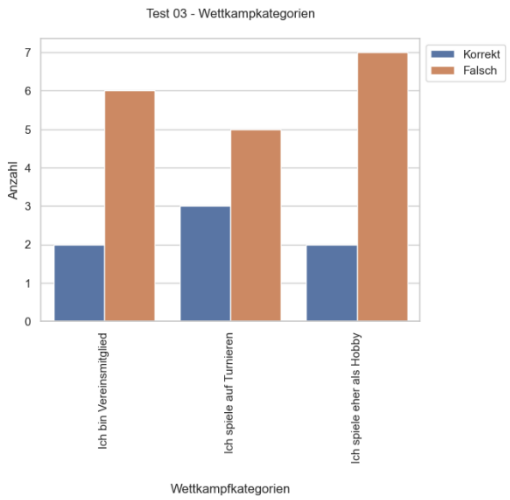


Abbildung 27: Wettkampfkategorien aus Test 03, Mehrfachnennung möglich.

Abbildung 27 zeigt verhältnismäßig mehr Hobbyspielende in der falsch urteilenden Gruppe, während in der anderen Gruppe die meisten auf Turnieren spielen.

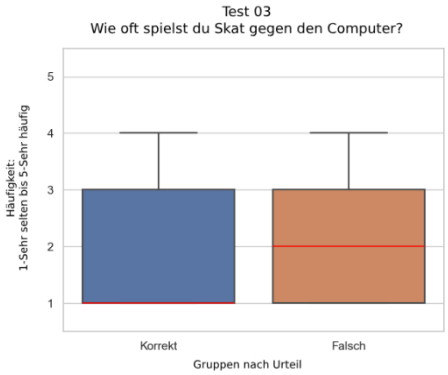


Abbildung 28: Erfahrung mit Bots aus Test 03.

In Abbildung 28 hat die Erfahrung mit Bots ein Tief erreicht: Der Median der Gruppe „Korrekt“ liegt bei eins. Damit ist sie zum ersten Mal weniger erfahren als die Gruppe „Falsch“.

Insgesamt hat die falsch urteilende Gruppe zum dritten Mal die besseren Skatspielenden und auch die Hobbyspielenden waren in dieser Gruppe verhältnismäßig am meisten vertreten. Zum ersten Mal sind die Gruppen nicht gleich unerfahren im Umgang mit Bots, die Gruppe „Korrekt“ ist im Median einen Punkt unter der anderen.

4.4. Test 04 – Skatfreunde unter sich

Im finalen Test ist eine Runde aus drei Fox KIs zu sehen. Vorhand wird mit 18 AS und sagt Karo an. Gezeigt wird die Perspektive von Hinterhand. Das Spiel wird mit 74 Augen vom AS gewonnen.

Glaubwürdigkeit

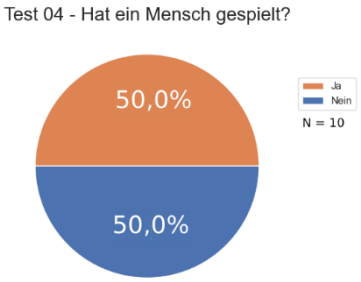


Abbildung 29: Testurteile in Test 04, korrektes Urteil „Nein“ in blau

Die Abbildung 29 zeigt ein uneiniges Urteil. Eine Hälfte der Befragten vermutet einen Menschen und die andere Hälfte einen Bot. Damit ist die Fox-KI weniger glaubwürdig als der menschliche Experte, was sie zum unglaubwürdigsten Testsubjekt in der Reihe macht.

Abbildung 30: Wahrgenommenes Fähigkeitsniveau in Test 04.

In Abbildung 30 haben Befragte das Fähigkeitsniveau im Video bewertet. Der Mensch erhält von den meisten eine zwei von fünf für die Partie, während der Bot am häufigsten um einen Punkt schlechter ist. Im Median werden beide Identitäten gleich gut bewertet. Mensch und Bot erhalten jeweils zwei Punkte.

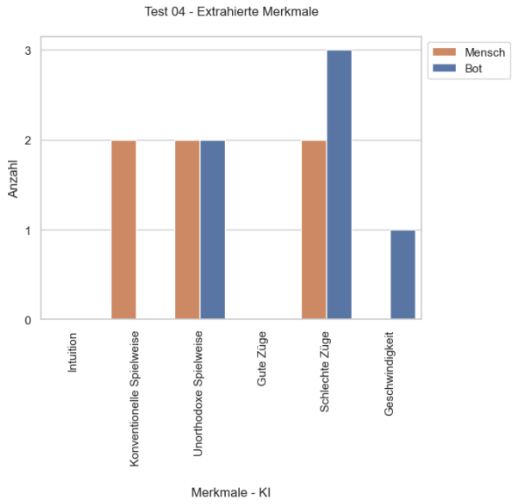


Abbildung 31: Merkmale aus Testurteilen von Test 04, mehrfaches Vorkommen möglich.

Abbildung 31 zeigt „Schlechte Züge“ als häufigste Begründung für die Urteile. Ein Kreuz Ass wurde überstürzt gespielt und verloren. Entsprechend ist „Unorthodoxe Spielweise“ an zweiter Stelle. Befragte ordnen dieses Merkmal beiden Identitäten ähnlich oft zu.

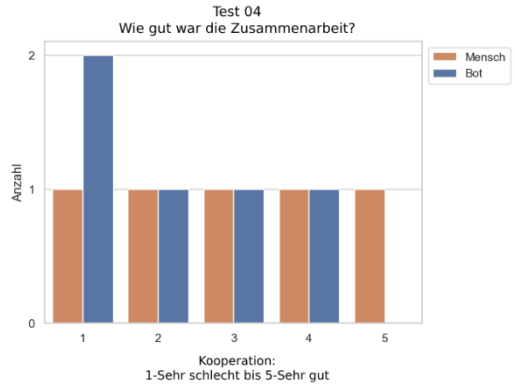


Abbildung 32: Wahrgenommene Kooperation in Test 04.

Über die Kooperation gibt es geteilte Meinungen. In Abbildung 32 ist zu sehen, dass die Werte über die Skala verstreut sind. Der Mensch erhält drei und der Bot zwei von fünf Punkten.

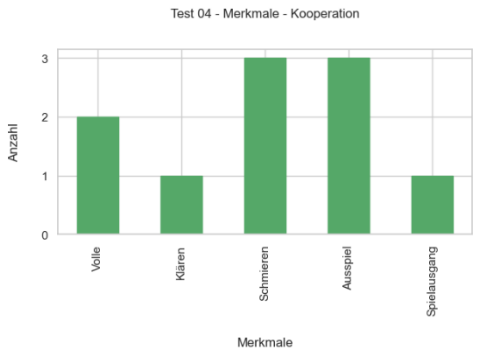


Abbildung 33: Merkmale für Kooperation in Test 04, N = 6, mehrfaches Vorkommen möglich.

Sechs von zehn Testteilnehmenden erklärten ihr Urteil über die Kooperation im Test. Abbildung 33 zeigt „Schmieren“ und „Ausspiel“ als häufigste Themen. Wie zuvor beschrieben wurde ein Kreuz Ass verfrüht ausgespielt, als man von Herz gewechselt hat.

Das Teammitglied im Video hat in Herz weder Ass noch Zehn ausgespielt, was als „Schnippeln“ gesehen wird. Diese Aktion wurde unter dem Merkmal „Schmieren“ verzeichnet. Die Herz-Zehn wird von vielen Befragten richtigerweise als gedrückt vermutet, weswegen sie gerne weiter auf Herz geblieben wären, um die Farbe zu klären und das Ass einzuholen.

Zum Schluss hält sich das Urteil in Test 04 die Waage. Aufgrund des zuvor beschriebenen Zuges entschieden sich Befragte zu einer Hälfte für die menschliche Identität und zur anderen Hälfte für den Bot.

Ein letztes Mal folgt nun der demografische Teil.

Demografie

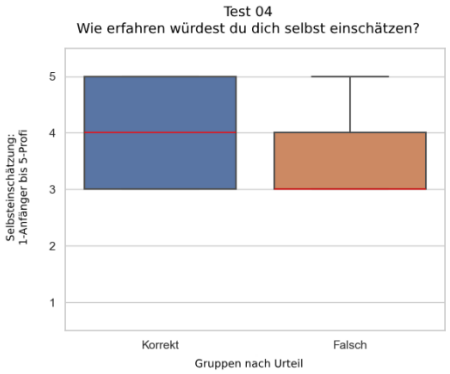


Abbildung 34: Selbsteinschätzung der Testpersonen aus Test 04.

Abbildung 34 zeigt ähnliche Mediane wie in den vorherigen Tests. Dieses Mal schätzt sich jedoch die Gruppe „Korrekt“ um einen Punkt stärker ein und nich umgekehrt. Allgemein sind in letzterer Gruppe mehr hohe Wertungen vertreten, was man an der größeren Box ablesen kann.

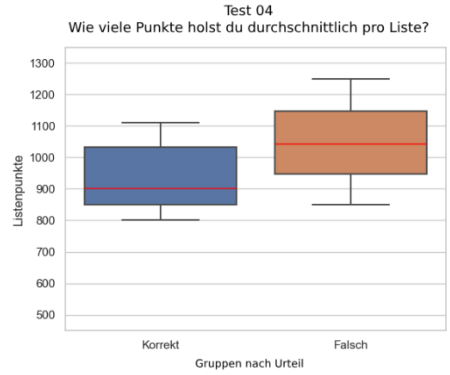


Abbildung 35: Listenpunkte aus Test 04, N = 8.

Während die Selbsteinschätzung umgekehrt wurde, ist in Abbildung 35 das Muster aus den vorherigen Tests wieder zu sehen. Die Gruppe „Falsch“ liegt in Listenpunkten vorne.

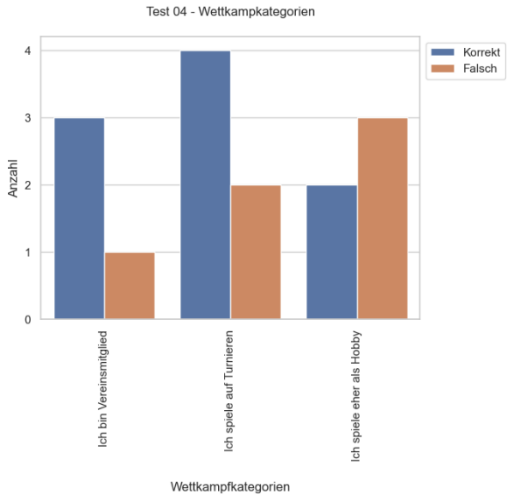


Abbildung 36: Wettkampfkategorien aus Test 04, Mehrfachnennung möglich.

Die Hobbyspielenden sind wieder am häufigsten in der Gruppe „Falsch“ vertreten. Abbildung 36 zeigt, dass die Gruppe „Korrekt“ kompetitiver ist.

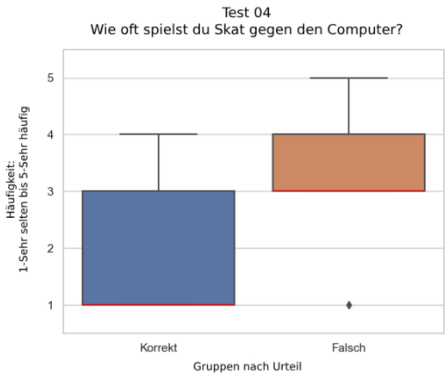


Abbildung 37: Erfahrung mit Bots aus Test 03.

In Abbildung 37 geht die Erfahrung mit Bots weit auseinander. Es existiert keine Überlappung zwischen den Boxen. Gruppe „Korrekt“ ist unerfahrener mit einem Median von eins, während Gruppe „Falsch“ mit Median drei durchschnittlich erfahren ist. Letzterer Median ist so hoch, dass der Wert eins ein Ausreißer wird.

Final weichen in Test 04 Selbsteinschätzung und angegebene Punkte voneinander ab. Ansonsten ist das Fazit ähnlich zu den anderen Tests: Gruppe „Falsch“ enthält die besseren Spieler und Hobbyspielende sind dort am meisten vertreten. Wie im vorhergegangenen Test geben öfter mit Bots spielende Befragte mehr falsche Antworten.

Die Ergebnisse der gesamten Testreihe und die Beantwortung der Forschungsfragen werden im nächsten Kapitel diskutiert.

5. Diskussion

Von allen getesteten Konstellationen war der Bot von Isar Interactive der glaubwürdigste. Er konnte 72,2% überzeugen. Die Fox KI aus Test 04 erwies sich als am wenigsten glaubwürdig, dicht gefolgt vom menschlichen Experten aus Test 01 (50% zu 53,2%).

Die Expertenrunde erhielt die höchsten Bewertungen für Fähigkeitsniveau und Kooperation. Knapp dahinter liegt der Bot von Isar. Test 02 hat die niedrigste Bewertung für Kooperation und Test 04 die niedrigste für Fähigkeit. Beide wahrgenommenen Eigenschaften unterscheiden sich meist nur um einen Punkt auf der Skala.

Über alle Tests hinweg wurde die unorthodoxe Spielweise des Testsubjekts am häufigsten als Indiz für das eigene Testurteil herangezogen. Am zweithäufigsten wurden schlechte Züge genannt, gefolgt von konventioneller Spielweise.

Bei der Bewertung der Kooperation war die Verwendung der Vollen am ausschlaggebendsten. Sie kamen nicht nur am häufigsten vor, sondern waren auch Bestandteil der meisten Stiche, die in den Freitexten zur Bewertung von Glaubwürdigkeit und Kooperation verwendet wurden.

Im demografischen Teil urteilen die besseren Spielenden öfter falsch als vergleichsweise schlechtere Spieler. Diese Beobachtung geht aus den im Mittel geringeren Listenpunkten und den größtenteils niedrigeren Selbsteinschätzungen in allen Tests hervor.

Die beschriebenen Unterschiede der demografischen Eigenschaften sind, ähnlich wie bei der Wahrnehmung, nicht sehr groß. Erlangte Listenpunkte unterscheiden sich im Mittel nur um ca. 50 Punkte, was in etwa ein bis zwei gewonnene Spiele sind. Auch die Selbsteinschätzung unterscheidet sich nur um einen Punkt auf der Skala. Eine Ausnahme stellt Test 04 dar, bei dem eine bessere Selbsteinschätzung häufiger mit einem korrekten Testurteil einherging.

Im Folgenden sollen nun die Ergebnisse bezüglich der einzelnen Forschungsfragen genauer diskutiert werden.

Forschungsfrage 1: Wie glaubwürdig sind eingesetzte Bots?

Innerhalb der durchgeführten Tests konnten die Bots viele Befragte von ihrer menschlichen Identität überzeugen. Der höchste Glaubwürdigkeitswert wurde mit 72,2% in Test 03 erreicht.

Betrachtet man die Gruppen genauer, so zeigt sich, dass die homogenen Testkonstellationen 01 und 04 am schlechtesten abschneiden. Die heterogenen Gruppen erreichen eine Glaubwürdigkeit von ca. 70%, während erstere nur um die 50% erreichen.

Dies legt die Vermutung nahe, dass Unterschiede zwischen den Teamspielenden mehr Möglichkeiten für ein differenziertes Testurteil bieten.

So könnte es beispielsweise sein, dass die Bots in Test 02 und Test 03 durch ihren menschlichen Mitspielenden glaubwürdiger erschienen als der Bot in Test 04, der mit einem anderen Bot zusammenspielte. Tatsächlich wurden Teammitglieder in beiden heterogenen Tests von Freitexturteilen zu Glaubwürdigkeit und Kooperation erwähnt.

Weiterführend könnte dies bedeuten, dass auch die Identität des AS ein Faktor ist. Damit wäre auch die Konstellation zwei Bots gegen einen Menschen nicht äquivalent zu Bot und Mensch gegen einen weiteren Bot, wie ursprünglich angenommen.

Dementsprechend könnte man in zukünftigen Arbeiten die Testkonstellationen stärker vereinheitlichen, sei es durch den Faktor Alleinspielender oder Mitspielender. Man könnte auch versuchen, die Vermutungen zu widerlegen, indem man eine weitere Testreihe mit ausschließlich homogenen Tests durchführt.

Forschungsfrage 2: Sind Fähigkeit und Kooperation wichtig, um als Mensch zu gelten?

Die Glaubwürdigkeit des einzigen Menschen fällt mit 53,2% relativ gering aus, obwohl er das höchste wahrgenommene Spielniveau hat. Lediglich die komplett aus Bots bestehende Runde in Test 04, die gleichzeitig das niedrigste wahrgenommene Spielniveau aufweist, hat eine geringere Glaubwürdigkeit.

Obwohl dem Spiel eines vermeintlichen Bots nie das Merkmal „Gute Züge“ zugeordnet wurde, unterscheiden sich die Fähigkeitsbewertungen von Mensch und Bot nur in zwei Tests. In diesen beiden Tests (01, 02) ist der Unterschied dazu noch gering. Lediglich ein Punkt trennt die beiden Gruppen in Test 01 und ein halber Punkt in Test 02.

Ähnlich sieht es bei der Kooperation aus: Die Wertungen unterscheiden sich nur in Tests 01 und 04, um jeweils einen Punkt.

	p-Wert	Normalverteilt	Homogene Varianz	Durchgeführter Test
Test 01	0,01487	Nein	entfällt	Mann-Whitney-U
Test 02	0,79889	Nein	entfällt	Mann-Whitney-U
Test 03	0,91477	Ja	Ja	t-Test
Test 04	0,75992	Ja	Ja	t-Test

Abbildung 38: Signifikanzwerte für wahrgenommene Spielstärke

	p-Wert	Normalverteilt	Homogene Varianz	Durchgeführter Test
Test 01	0,1355	Nein	entfällt	Mann-Whitney-U
Test 02	0,7137	Nein	entfällt	Mann-Whitney-U
Test 03	1,0000	Nein	entfällt	Mann-Whitney-U
Test 04	0,4082	Ja	Ja	t-Test

Abbildung 39: Signifikanzwerte für wahrgenommene Kooperation

Zur weiteren Analyse der Ergebnisse wurden statistische Tests durchgeführt, die aufzeigen sollen, ob sich die zentrale Tendenz der beiden Gruppen, „Korrekt“ und „Falsch“, signifikant voneinander unterscheidet. Die Ergebnisse sowie durchgeführte Tests sind in den Abbildungen 38 und 39 zu sehen.

Es ist zu erkennen, dass allein die Expertenrunde im Bereich wahrgenommene Spielstärke ein signifikantes Ergebnis gegen das verwendete Signifikanzniveau von 0,05 aufweist. Alle anderen Tests liefern nicht signifikante Ergebnisse. Da der signifikante Test auch gleichzeitig der mit den meisten Befragten ist, bietet sich wahrscheinlich eine Wiederholung der Testreihe an, bei der eine gleiche Anzahl an Antworten pro Test sichergestellt wird. Eine solche Gewährleistung war im zeitlichen Rahmen dieser Masterarbeit leider nicht möglich.

Abschließend suggerieren die mehrheitlich nicht signifikanten Ergebnisse, dass es keine Korrelation zwischen dem wahrgenommenen Fähigkeitsniveau des Testsubjekts und dessen Glaubwürdigkeit gibt, ähnlich wie (Hingston, 2009) es beobachtete und im Gegensatz zu (Laird & Duchi, 2000).

Für Kooperationsfähigkeit ist das Urteil nicht eindeutig, da sie in dieser Arbeit indirekt bewertet wurde. In den behandelten verwandten Arbeiten wurde stets direkt mit den Bots interagiert (Correia et al., 2016). Insofern könnte das Ergebnis der Studie bedeuten, dass die direkte Interaktion sich schlichtweg besser zur Beurteilung von Mensch-Maschinen-Kooperation eignet als indirekte Bewertung.

Forschungsfrage 3: Warum werden Bots als solche identifiziert oder warum nicht?

Insgesamt konnten sechs verschiedene Merkmale aus den Freitextantworten extrahiert werden, die von Testteilnehmenden zur Urteilsfindung herangezogen wurden:

– Intuition,
– Konventionelle Spielweise,
– Unorthodoxe Spielweise,
– Gute Züge,
– Schlechte Züge,
– Geschwindigkeit.

Die häufigsten davon waren:

– Unorthodoxe Spielweise,
– Schlechte Züge,
– Konventionelle Spielweise.

Leider weist keines dieser Merkmale eine hohe Erfolgsquote auf. So war „unorthodoxe Spielweise“ im ersten Test entscheidend für die meisten Urteile, die den Experten fälschlicherweise als Bot identifizierten. In den folgenden Tests war es dann gleichverteilt zwischen korrekt und falsch.

„Schlechte Züge“ wurden ebenfalls häufiger der falschen Testidentität zugeordnet, außer im letzten Test, wo sie am häufigsten ein richtiges Urteil bedingten. Am häufigsten wurden sie genutzt, um das Testsubjekt als Mensch zu identifizieren.

Bemerkenswert ist auch, dass das gegenteilige Merkmal „Gute Züge“ kein einziges Mal als Argument für einen Bot verwendet wurde. Ähnlich verhält es sich mit „konventionelle[r] Spielweise“, die den Bots nur einmal zugeschrieben wurde.

Die Teilnehmenden konstruieren folgendes Bild von Mensch und Bot: Ein Mensch hält sich an die Skat-Merksätze. Es werden die bekannten Konventionen im Gegenspiel befolgt, wobei jedoch offensichtliche Fehler gemacht werden. Ein Bot spielt weniger fehlerbehaftet, weicht aber oft von bekannter Spielweise ab, was manchmal funktioniert, aber niemals großartig ist.

In den Tests wurde eine normierte Ausspielgeschwindigkeit verwendet, da Geschwindigkeit in verwandten Tests häufig zur Identifikation verwendet wurde. Leider taucht der Faktor Geschwindigkeit trotz dieser Maßnahme in den Testmerkmalen auf.

Grund dafür ist, dass die Testteilnehmenden zu Beginn des Fragebogens nicht über die manipulierte Darstellung des Spielablaufs informiert wurden. Dieser Fehler wurde bereits nach dem ersten Test bemerkt, aber dennoch beibehalten, um gleiche Bedingungen zwischen den Tests zu gewährleisten.

Schlussendlich bleibt die Frage, warum die Testteilnehmenden so große Schwierigkeiten hatten, im Test korrekt zu urteilen. Dazu werfen wir einen Blick auf die demografischen Fragen.

Zwei der demografischen Fragen, genauer gesagt Selbsteinschätzung und Listenpunkte, sollten das Niveau der Befragten bestimmen. Hier konnten die besseren Spielenden Bots schlechter von Menschen unterscheiden als die schlechteren. Sowohl bei Selbsteinschätzung als auch in den Listenpunkten waren die Werte der falsch urteilenden Gruppe höher.

	p-Wert	Normalverteilt	Homogene Varianz	Durchgeführter Test
Test 01	0,8530	Nein	entfällt	Mann-Whitney-U
Test 02	0,4623	Nein	entfällt	Mann-Whitney-U
Test 03	0,5604	Nein	entfällt	Mann-Whitney-U
Test 04	0,5708	Nein	entfällt	Mann-Whitney-U

Abbildung 40: Signifikanzwerte der Selbsteinschätzungen

	p-Wert	Normalverteilt	Homogene Varianz	Durchgeführter Test
Test 01	0,7458	Ja	Ja	t-Test
Test 02	0,0600	Nein	entfällt	Mann-Whitney-U
Test 03	0,5787	Ja	Ja	t-Test
Test 04	0,3780	Ja	Ja	t-Test

Abbildung 41: Signifikanzwerte der Listenpunkte

Wie man in den Abbildungen 40 und 41 sieht, ist dieser Unterschied jedoch nicht groß genug, um signifikant zu sein. Es bleibt also weiterhin zu klären, ob höhere Spielstärke hilfreich ist, um im Gegenspiel erfolgreich zwischen Mensch und Bot zu unterscheiden.

Ein ähnliches Bild zeigt sich bei der Erfahrung mit Bots, wo die falsch urteilenden Befragten geringfügig weniger Erfahrung angeben als die korrekt urteilenden. Mit zwei von fünf Punkten Differenz im Mittel, ist der Unterschied bei Test 04 am größten. Da dieser Test jedoch die geringste Anzahl an Antworten hat, ist dies wenig aussagekräftig. In Test 01, welcher die meisten Antworten bekam, sind die Erfahrungen fast gleichverteilt.

	p-Wert	Normalverteilt	Homogene Varianz	Durchgeführter Test
Test 01	0,8919	Nein	entfällt	Mann-Whitney-U
Test 02	0,3011	Nein	entfällt	Mann-Whitney-U
Test 03	0,7176	Nein	entfällt	Mann-Whitney-U
Test 04	0,2268	Ja	Ja	t-Test

Abbildung 42: Signifikanzwerte der Erfahrungen mit Bots

Dementsprechend ist in Abbildung 42 zu erkennen, dass es in keinem Test einen signifikanten Unterschied zwischen den beiden Gruppen gab.

Da keine der erfassten demografischen Eigenschaften einen signifikanten Zusammenhang zum Testurteil aufweist, lässt sich also vermuten, dass die Befragten Erfahrungen außerhalb von Skat in ihr Urteil einfließen lassen. Ein Beispiel dafür könnte ihre eigene Meinung gegenüber der Kompetenz bzw. Spielstärke von KI sein.

Es ist nicht bekannt, ob Befragte wissen, dass KI beim Skat nicht viel stärker ist als der Durchschnitt der Testteilnehmenden. Ebenso gut könnten sie den Eindruck einer Überlegenheit von KI aus der aktuellen Berichterstattung bzw. dem Diskurs über Technologien wie ChatGpt gewonnen haben.

Dass die Meinungen über die Fähigkeiten von Skat-Bots unter den Befragten auseinandergehen müssen, wird daran deutlich, dass sie bei den gleichen Stichen unterschiedliche Urteile fällen. So in jedem einzelnen Test geschehen.

Schlussendlich scheint Erfahrung mit Skat keine Hilfe bei der Erkennung von Bots zu sein. Gleiches scheint für Erfahrung mit Skat-Bots zu gelten. Daher ist es wahrscheinlich, dass die Testenden aufgrund ihrer eigenen Meinung zur Technologie KI entscheiden. Inwieweit dies einen Einfluss auf den Test hat, könnte in weiterer Arbeit erforscht werden.

Forschungsfrage 4: Welche Merkmale stehen für gute Kooperation?

Für die Kooperation wurden acht Merkmale gebildet:
- Volle,
- Klären,
- Nachspielen,
- Stellungsspiel,
- Schmieren,
- Ausspiel,
- Trumpf,
- Spielausgang.

Das wichtigste Merkmal ist die Verwendung der Vollen. Wann man diese Karten ausspielt oder schmiert, ist laut den Befragten entscheidend. Es ist wichtig zu wissen, wann man mit ihnen noch Stiche machen kann oder wann man Volle schmieren muss, um sie nicht zu verlieren. Der Verlust von vollen Karten wurde von den Testenden immer wieder bemängelt.

Entsprechend bezogen sich Urteilsbegründungen meist auf Züge, bei denen volle Karten im Stich waren. Beispielsweise wurde in Test 01 häufig das Ausspielen der Herz 10 kritisiert. Dort sollte laut Konvention der Kreuz-König gespielt werden.

Auch Ausspiel und Klärung wurden sehr oft genannt. Sie seien am wichtigsten für Kommunikation zwischen den GS. Die Klärung einer Farbe gebe Teammitgliedern wichtige Informationen für das Ausspielen (von Vollen). In Test 02 z.B. hätte der entscheidende Spielfehler nach Meinung vieler Befragter durch korrektes Ausspiel vermieden werden können.

Ein unerwünschtes Merkmal aus der Befragung ist der „Spielausgang“. Für einige war es ausreichend, ob ein Spiel gewonnen oder verloren wurde, um ihre Bewertung zu vergeben.

Obwohl in Test 02 nur ein einziger Fehler beanstandet wurde, erhielt der Bot dafür sehr viele niedrige Bewertungen, weil Schneider verloren wurde. Umgekehrt erhielt der Bot in Test 03 viele hohe Bewertungen, weil das Gegenspiel gewonnen wurde. Die hohen Bewertungen gab es auch trotz der Anmerkungen, dass eigentlich keine besonderen Züge nötig gewesen wären.

Es ist also zu empfehlen, in allen Testszenarien einen gleichwertigen Spielausgang zu zeigen, da Testende davon beeinflusst werden.

Für Entwickelnde von Skat-Bots bieten die extrahierten Merkmale Ansätze für die Erstellung einer eigenen kooperativen KI. Eine solche KI könnte sich besonders auf die in den Tests aufgetretenen Merkmale fokussieren und versuchen, in diesen Bereichen besonders gut zu sein. So könnte Frust für menschliche Teammitglieder minimiert und damit der Spielspaß im Gegenspiel erhöht werden.

Das Potenzial einer separaten KI für das Gegenspiel wird deutlich, wenn man sich noch einmal die Unfähigkeit der gängigen Monte-Carlo-Methode vor Augen führt, klärende bzw. informationsgewinnende Züge zu spielen (Kupferschmid, 2003). Die Wichtigkeit solcher Züge wurde nicht nur durch das wiederholte Auftreten des Merkmals in den Tests unterstrichen, sondern auch durch Kritik am Spiel der Testsubjekte.

Dementsprechend könnte es gar sein, dass regelbasierte KI für das Gegenspiel besser geeignet ist als andere spielstärkere Methoden. Vorangegangene Forschung zeigt dafür ebenfalls Potenzial (Correia et al., 2016).

6. Fazit

Im Rahmen dieser Arbeit wurden erstmals Skat-Bots auf ihre Glaubwürdigkeit als menschliche Mitspielende untersucht. In vier separaten Tests wurde den Mitgliedern des Skatfreunde Discord-Servers jeweils eine separate Partie Skat gezeigt. Aufgrund der gesehenen Partie sollten sie dann beurteilen, ob diese von einem Menschen gespielt wurde. Die Befragten beantworteten asynchron einen Fragebogen, in den die Aufnahmen eingebettet waren.

Testsubjekte waren ein menschlicher Experte und drei Bots aus kommerziellen Produkten. Diese wurden in unterschiedlichen Konstellationen als Gegenspielende gezeigt. In den Tests konnte der Bot von Isar Interactive 73% der Befragten davon überzeugen, dass er ein menschlicher Skatspielender war. Generell hatte kein Testsubjekt eine Glaubwürdigkeit unter 50%. Des Weiteren fiel auf, dass homogene Testkonstellationen niedrigere Glaubwürdigkeitswerte erzielten als heterogene.

Die Testurteile waren in Freitexten zu begründen, so dass aus den Antworten Merkmale für menschliche Spielweise bzw. Spielweise von Bots gewonnen werden konnte. Menschliche Spielende halten sich den Antworten zufolge streng an Skat-Merksätze und andere bekannte Konventionen. Ihnen werden mehr Fehler zugeschrieben als den Bots. Ein Bot hingegen spiele oft unorthodox, aber fehlerfreier.

Die Wahrnehmung als Mensch bzw. Bot hatte in dieser Testreihe keinen Einfluss auf die Bewertung von Spielstärke und Kooperationsfähigkeit. Die Befragten bewerteten vermeintliche Bots in der Regel genauso gut wie vermeintliche Menschen.

Bei den Beurteilenden selbst konnte hohe Spielstärke nicht beim Test helfen. Bessere Spielende antworteten häufiger falsch. Befragte konnten auch kaum auf nennenswerte Erfahrungen mit Bots zurückgreifen, die sie für ihr Urteil hätten nutzen können. Stattdessen müssen sie sich auf ihre eigene Meinung gegenüber der KI verlassen, welche im Test nicht erfasst wurde. Diese Vermutung wird dadurch gestützt, dass Befragte stets ein oder zwei Stiche aus einem Video hervorhoben, um dann anhand ein und desselben Stiches unterschiedliche Urteile zu fällen.

Zuletzt wurden auch Merkmale für Kooperation aus den Freitextantworten gesammelt. Hier wurde am häufigsten die Verwendung der Vollen in Kombination mit Schmieren und Ausspiel genannt. Das wichtigste Merkmal zur Kommunikation war Klärung. Gerade klärende Züge werden von den bekannten KI-Ansätzen oft vermieden.

Entsprechend bietet es sich für zukünftige Forschung an Skat-KI bzw. Skat-Bots an, eigene Module für das Gegenspiel zu entwickeln, die sich an bekannte Konventionen unter den Skatspielenden halten und auch kommunikative Züge zulassen.

Für den Skat-Turing-Test selbst bietet es sich an, noch mehr Parameter der gezeigten Partien zu normieren. Beispiele dafür wären:

- Kartenverteilung, da in dieser Ausführung kein Einfluss darauf genommen werden konnte,
- Testkonstellationen,
- Spielausgang.

Des Weiteren könnte man den Test mit den Spezialspielen Null und Grand wiederholen. Gerade Null unterscheidet sich massiv von den im Test gezeigten Farbspielen.

Für Skat-Bots ist der Test ein Erfolg, da sie bereits zu zwei Dritteln als Menschen überzeugen können. Allerdings bleibt festzustellen, ob diese Ergebnisse mit anderen Testteilnehmenden reproduzierbar sind. Die befragte Community zeichnete sich vor allem durch wettbewerbsorientierte Spielende aus, die Skat hauptsächlich gegen andere Menschen spielen.

Ein neuer Test könnte vor allem mit KI-erfahrenen Befragten durchgeführt werden. Dies könnte dann auch die Vermutung bestätigen oder widerlegen, dass die Befragten im Rahmen der durchgeführten Tests vor allem ihre eigene Meinung gegenüber KI für ihr Urteil genutzt haben.

7. Literaturverzeichnis

Ashktorab, Z., Liao, Q. V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2020). Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 96:1-96:20. https://doi.org/10.1145/3415167

Baier, H., Sattaur, A., Powley, E. J., Devlin, S., Rollason, J., & Cowling, P. I. (2019). Emulating Human Play in a Leading Mobile Card Game. *IEEE Transactions on Games*, *11*(4), 386–395. https://doi.org/10.1109/TG.2018.2835764

Buro, M. (2007, Oktober 16). *ISS - Home*. https://skatgame.net/iss/

Buro, M., Long, J. R., Furtak, T., & Sturtevant, N. (2009). Improving state evaluation, inference, and search in trick-based card games. *Twenty-First International Joint Conference on Artificial Intelligence*.

Correia, F., Alves-Oliveira, P., Maia, N., Ribeiro, T., Petisca, S., Melo, F. S., & Paiva, A. (2016). Just follow the suit! Trust in human-robot interactions during card game playing. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 507–512. https://doi.org/10.1109/ROMAN.2016.7745165

Cowling, P. I., Devlin, S., Powley, E. J., Whitehouse, D., & Rollason, J. (2015). Player Preference and Style in a Leading Mobile Card Game. *IEEE Transactions on Computational Intelligence and AI in Games*, *7*(3), 233–242. IEEE Transactions on Computational Intelligence and AI in Games. https://doi.org/10.1109/TCIAIG.2014.2357174

Devlin, S., Anspoka, A., Sephton, N., Cowling, P., & Rollason, J. (2016). Combining Gameplay Data with Monte Carlo Tree Search to Emulate Human Play. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *12*(1), Article 1. https://doi.org/10.1609/aiide.v12i1.12858

Devlin, S., Georgescu, R., Momennejad, I., Rzepecki, J., Zuniga, E., Costello, G., Leroy, G., Shaw, A., & Hofmann, K. (2021). Navigation Turing Test (NTT): Learning to Evaluate Human-Like Navigation. *Proceedings of the 38th International Conference on Machine Learning*, 2644–2653. https://proceedings.mlr.press/v139/devlin21a.html

Discord Inc. (2024, Januar 18). *Discord Anfängerleitfaden*. Discord. https://support.discord.com/hc/de/articles/360045138571-Discord-Anf%C3%A4ngerleitfaden

Edelkamp, S. (2020). Representing and Reducing Uncertainty for Enumerating the Belief Space to Improve Endgame Play in Skat. In *ECAI 2020* (S. 395–402). IOS Press. https://doi.org/10.3233/FAIA200118

Edelkamp, S. (2021a). Challenging Human Supremacy in Skat. *Proceedings of the International Symposium on Combinatorial Search*, *10*(1), 52–60. https://doi.org/10.1609/socs.v10i1.18502

Edelkamp, S. (2021b). *On the Power of Refined Skat Selection* (arXiv:2104.02997). arXiv. https://doi.org/10.48550/arXiv.2104.02997

Frank, I., & Basin, D. (1998). Search in games with incomplete information: A case study using Bridge card play. *Artificial Intelligence*, *100*(1–2), 87–123. https://doi.org/10.1016/S0004-3702(97)00082-9

Gao, Y., Liu, F., Wang, L., Lian, Z., Wang, W., Li, S., Wang, X., Zeng, X., Wang, R., Wang, J., Fu, Q., Wei, Y., Huang, L., & Liu, W. (2022). *Towards Effective and Interpretable Human-AI Collaboration in MOBA Games*. https://openreview.net/forum?id=2nJdh_C-UWe

Gerhardt, G. (2004, Dezember 4). *Die XSkat Home Page*. https://web.archive.org/web/20041204051251/http://www.xskat.de/index.html

Ginsberg, M. L. (1999). GIB: Steps toward an expert-level bridge-playing program. *IJCAI*, 584–593.

Gößl, R. (2019). *Der Skatfuchs – Gewinnen im Skatspiel mit Mathematische Methoden*. Eigenverlag. https://www.skatfuchs.eu/skatfuchs.htm

Gößl, R. (2020). *Skatfuchs*. https://www.skatfuchs.eu/skatfuchs.htm

Hingston, P. (2009). A Turing Test for Computer Game Bots. *IEEE Transactions on Computational Intelligence and AI in Games*, *1*(3), 169–186. IEEE Transactions on Computational Intelligence and AI in Games. https://doi.org/10.1109/TCIAIG.2009.2032534

Hingston, P. (2010). *A new design for a Turing Test for Bots*. 345–350. https://doi.org/10.1109/ITW.2010.5593336

Isar Interactive GmbH & Co Kg. (2014, Juli 11). *Www.skat-spiel.de*. https://www.skat-spiel.de

Keller, T., & Kupferschmid, S. (2008). Automatic Bidding for the Game of Skat. In A. R. Dengel, K. Berns, T. M. Breuel, F. Bomarius, & T. R. Roth-Berghofer (Hrsg.), *KI 2008: Advances in Artificial Intelligence* (Bd. 5243, S. 95–102). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-85845-4_12

Khalifa, A., Isaksen, A., Togelius, J., & Nealen, A. (2016). *Modifying MCTS for human-like general video game playing*. https://www.um.edu.mt/library/oar/handle/123456789/81986

Kupferschmid, S. (2003). *Entwicklung eines Double-Dummy Skat Solvers mit einer Anwendung für verdeckte Skatspiele* [Master's Thesis]. Albert-Ludwigs-Universität Freiburg.

Kupferschmid, S., & Helmert, M. (2007). A Skat Player Based on Monte-Carlo Simulation. In H. J. van den Herik, P. Ciancarini, & H. H. L. M. Donkers (Hrsg.), *Computers and Games* (Bd. 4630, S. 135–147). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75538-8_12

Laird, J. E., & Duchi, J. C. (2000). Creating human-like synthetic characters with multiple skill levels: A case study using the soar quakebot. *AAAI 2000 Fall Symposium Series: Simulating Human Agents*, *1001*, 48109-2110.3.

Lidén, L. (2003). Artificial stupidity: The art of intentional mistakes. *AI game programming wisdom*, *2*(5), 41–48.

Long, J. R. (2011). *Search, inference and opponent modelling in an expert-caliber Skat player*.

Lopez, S. A. (2005, August 21). Intelligent mistakes. *Chess News*. https://en.chessbase.com/post/intelligent-mistakes

Merritt, T. R., Tan, K. B., Ong, C., Thomas, A., Chuah, T. L., & McGee, K. (2011). Are artificial team-mates scapegoats in computer games. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 685–688. https://doi.org/10.1145/1958824.1958945

Milani, S., Juliani, A., Momennejad, I., Georgescu, R., Rzpecki, J., Shaw, A., Costello, G., Fang, F., Devlin, S., & Hofmann, K. (2023). Navigates Like Me: Understanding How People Evaluate Human-Like AI in Video Games. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3544548.3581348

Newborn, M. (1997). Deep Blue and Garry Kasparov in Philadelphia. In M. Newborn (Hrsg.), *Kasparov versus Deep Blue: Computer Chess Comes of Age* (S. 235–278). Springer. https://doi.org/10.1007/978-1-4612-2260-6_9

Niklaus, J., Alberti, M., Pondenkandath, V., Ingold, R., & Liwicki, M. (2019). *Survey of Artificial Intelligence for Card Games and Its Application to the Swiss Game Jass* (arXiv:1906.04439). arXiv. http://arxiv.org/abs/1906.04439

OpenAI, Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H. P. d O., Raiman, J., Salimans, T., … Zhang, S. (2019). *Dota 2 with Large Scale Deep Reinforcement Learning* (arXiv:1912.06680). arXiv. https://doi.org/10.48550/arXiv.1912.06680

Ortega, J., Shaker, N., Togelius, J., & Yannakakis, G. N. (2013). Imitating human playing styles in Super Mario Bros. *Entertainment Computing*, *4*(2), 93–104. https://doi.org/10.1016/j.entcom.2012.10.001

Rebstock, D., Solinas, C., & Buro, M. (2019). *Learning Policies from Human Data for Skat* (arXiv:1905.10907). arXiv. https://doi.org/10.48550/arXiv.1905.10907

Schäfer, D. (2016, Oktober 7). *Skat lernen — YouTube*. https://www.youtube.com/channel/UC5pFGjXydPM_5Dblr-n4yQQ

Schäfer, J., Buro, M., & Hartmann, K. (2008). *The UCT algorithm applied to games with imperfect information*. Diploma thesis, Otto-Von-Guericke Univ. Magdeburg, Germany.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), Article 7587. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), Article 7676. https://doi.org/10.1038/nature24270

Siu, H. C., Pena, J. D., Chen, E., Zhou, Y., Lopez, V. J., Palko, K., Chang, K. C., & Allen, R. E. (2021). *Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi* (arXiv:2107.07630). arXiv. https://doi.org/10.48550/arXiv.2107.07630

Solinas, C., Rebstock, D., & Buro, M. (2019). Improving Search with Supervised Learning in Trick-Based Card Games. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), Article 01. https://doi.org/10.1609/aaai.v33i01.33011158

Soni, B., & Hingston, P. (2008). Bots trained to play like a human are more fun. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 363–369. https://doi.org/10.1109/IJCNN.2008.4633818

StammtischGames GmbH & Co. KG. (2015, April 25). *Skat am Stammtisch – Apps bei Google Play*. https://play.google.com/store/apps/details?id=skat.am.stammtisch.frei&hl=de

Togelius, J., Yannakakis, G. N., Karakovskiy, S., & Shaker, N. (2012). Assessing Believability. In P. Hingston (Hrsg.), *Believable Bots: Can Computers Play Like People?* (S. 215–230). Springer. https://doi.org/10.1007/978-3-642-32323-2_9

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., … Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(7782), Article 7782. https://doi.org/10.1038/s41586-019-1724-z

Wehbe, R. R., Lank, E., & Nacke, L. E. (2017). Left Them 4 Dead: Perception of Humans versus Non-Player Character Teammates in Cooperative Gameplay. *Proceedings of the 2017 Conference on Designing Interactive Systems*, 403–415. https://doi.org/10.1145/3064663.3064712

West, M. (2009, März 18). Intelligent Mistakes: How to Incorporate Stupidity Into Your AI Code. *Game Developer*. https://www.gamedeveloper.com/programming/intelligent-mistakes-how-to-incorporate-stupidity-into-your-ai-code

Yannakakis, G. N., & Togelius, J. (2018). AI Methods. In G. N. Yannakakis & J. Togelius (Hrsg.), *Artificial Intelligence and Games* (S. 29–88). Springer International Publishing. https://doi.org/10.1007/978-3-319-63519-4_2

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). „An Ideal Human“: Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 246:1-246:25. https://doi.org/10.1145/3432945

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Master-Thesis mit dem Titel:

Der Skat-KI Turing-Test

Ermittlung der Eigenschaften glaubhafter Mitspieler-KI in Skat

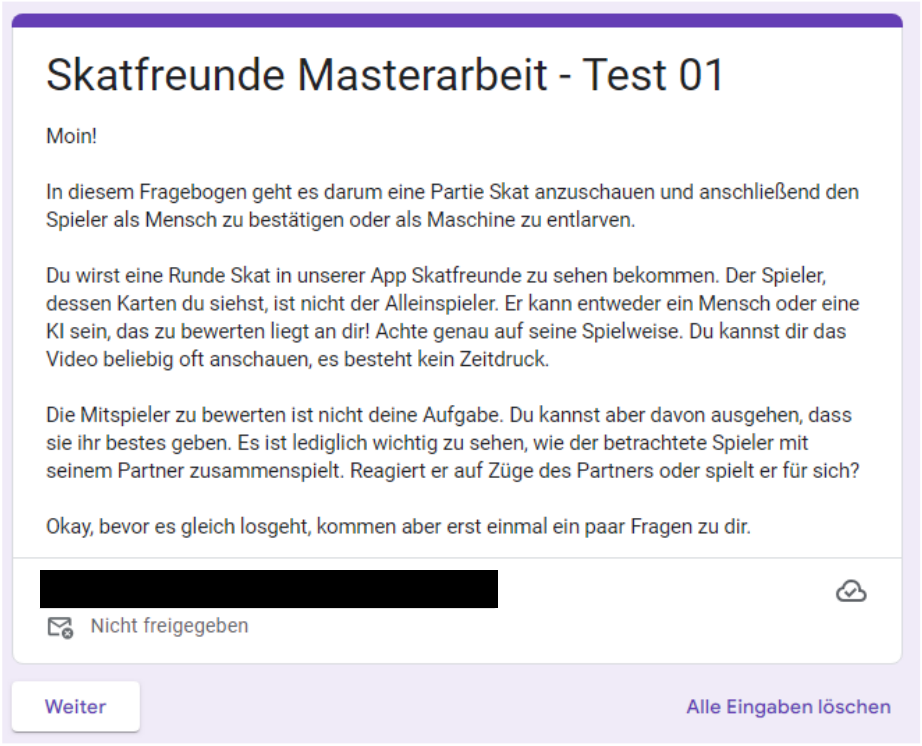
selbständig und nur mit den angegebenen Hilfsmitteln verfasst habe.
Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

[redacted], [redacted] .....

Datum, Unterschrift

Anhang

Fragebogen des ersten Tests; Grundlage für alle weiteren Tests



Skatfreunde Masterarbeit - Test 01

Moin!

In diesem Fragebogen geht es darum eine Partie Skat anzuschauen und anschließend den Spieler als Mensch zu bestätigen oder als Maschine zu entlarven.

Du wirst eine Runde Skat in unserer App Skatfreunde zu sehen bekommen. Der Spieler, dessen Karten du siehst, ist nicht der Alleinspieler. Er kann entweder ein Mensch oder eine KI sein, das zu bewerten liegt an dir! Achte genau auf seine Spielweise. Du kannst dir das Video beliebig oft anschauen, es besteht kein Zeitdruck.

Die Mitspieler zu bewerten ist nicht deine Aufgabe. Du kannst aber davon ausgehen, dass sie ihr bestes geben. Es ist lediglich wichtig zu sehen, wie der betrachtete Spieler mit seinem Partner zusammenspielt. Reagiert er auf Züge des Partners oder spielt er für sich?

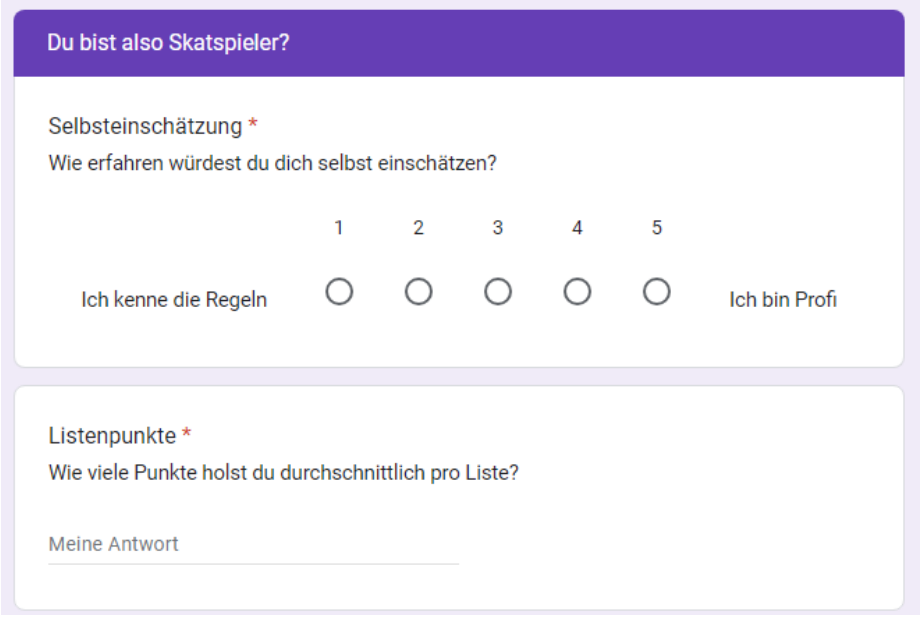
Okay, bevor es gleich losgeht, kommen aber erst einmal ein paar Fragen zu dir.

[redacted]

Nicht freigegeben

Weiter

Alle Eingaben löschen



Du bist also Skatspieler?

Selbsteinschätzung *

Wie erfahren würdest du dich selbst einschätzen?

	1	2	3	4	5	
Ich kenne die Regeln	○	○	○	○	○	Ich bin Profi

Listenpunkte *

Wie viele Punkte holst du durchschnittlich pro Liste?

Meine Antwort

Wettkampf

Bist du in einem Skat-Verein oder nimmst an Turnieren teil?

- [ ] Ich bin Vereinsmitglied
- [ ] Ich spiele auf Turnieren
- [ ] Ich spiele eher als Hobby

Erfahrungen mit Bots *

Wie oft spielst du Skat gegen den Computer?

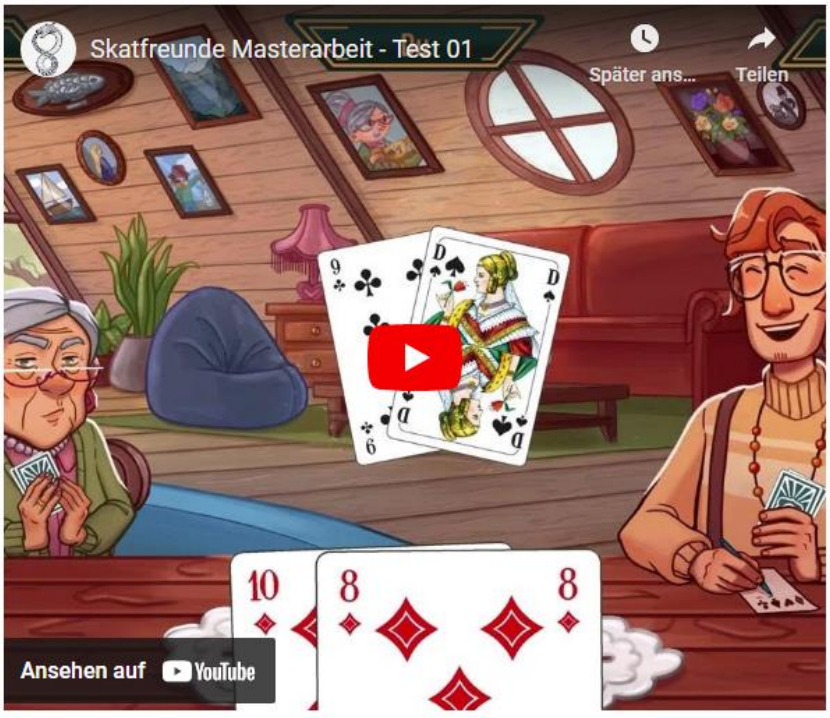
	1	2	3	4	5	
Ich spiele am meisten mit anderen Personen	○	○	○	○	○	Ich spiele am meisten gegen den Computer

Der Test

Reizgebote:
Mittelhand, Rechts sitzend -> hat sofort gepasst
Hinterhand, Spieler -> hat sofort gepasst

Vorhand, Links sitzend, bekommt das Spiel mit 18. Der Skat wird aufgenommen und es wird KREUZ gespielt.

Das Testvideo



Wurde die Partie von einem Menschen gespielt? *

- ◯ Ja
- ◯ Nein

Begründe bitte deine Wahl! *

Meine Antwort

Wie hoch schätzt du das Niveau des Spielers ein? *

	1	2	3	4	5	
Blutiger Anfänger	◯	◯	◯	◯	◯	Eiskalter Profi

Wie gut haben die beiden Gegenspieler zusammengearbeitet? *

	1	2	3	4	5	
Jeder hat für sich gespielt	◯	◯	◯	◯	◯	Sie waren ein eingespieltes Team

Kannst du sagen, warum die Gegenspieler gut bzw. schlecht gespielt haben?

Meine Antwort

Ende

Vielen Dank für deine Teilnahme!

Dies war der erste Fragebogen einer Serie. In den nächsten Tagen wird es regelmäßig so einen auf dem Discord geben. Ich würde mich freuen, wenn du auch das nächste Mal wieder dabei bist.

Ich weiß, es brennt dir bestimmt in den Fingern das Video zu diskutieren, bitte warte jedoch ein wenig, bevor du deine Analyse mit den anderen teilst.