LERNEN DURCH DIFFERENZ. ZUR LOGISCH-MATHEMATISCHEN STRUKTUR MASCHINELLEN LERNENS.

Peter Klimczak[1], *Günther Wirsching*[2], *Matthias Wolff*[1]
[1]*BTU Cottbus-Senftenberg,* [2]*KU Eichstätt-Ingolstadt*
guenther.wirsching@ku.de

Kurzfassung: Diese theoretische Untersuchung hat das Ziel, eine für maschinelles Lernen geeignete logisch-mathematische Formulierung des operanten Konditionierens nach B. F. Skinner zu finden. Dazu wählen wir zu jedem Lernreiz eine Differenzvariable, wobei wir unter Differenz eine Wahrnehmung von Unterschieden auf Seiten des Lernenden zwischen dem Zustand der Welt vor und nach seinem Verhalten verstehen. Da ein Lernreiz nicht vorhanden oder vorhanden und im letzteren Fall negativ oder positiv sein kann, besteht der Wertebereich einer Differenzvariablen im einfachsten Fall aus drei möglichen Werten: $+1$, 0 und -1. Abschließend stellen wir eine Verbindung zum Stevensschen Potenzgesetz her und skizzieren eine quantenlogische Modellierung.

1 Operantes Konditionieren

In der tradierten Vorstellung des operanten Konditionierens [1] geht man von einem Lernenden und einem Lehrenden aus. Letzterer intendiert ein bestimmtes Lernendenverhalten und versucht dieses durch Reizänderung zu erreichen. Man unterscheidet zwischen positiven (appetitiven) und negativen (aversiven) Reizen, Reizzustände also, die der Lernende erfahren will und deshalb positiv bewertet und Reizzustände, die er nicht erfahren will und daher negativ bewertet. Dabei wird nicht die bloße Anwesenheit oder Abwesenheit eines Reizes als entscheidend angesehen, sondern der Wechsel von der Anwesenheit hin zur Abwesenheit eines Reizes und vice versa. Dementsprechend ergeben sich vier Möglichkeiten, zwei Arten der Belohnung und zwei Arten der Bestrafung. Eine Belohnung kommt entweder durch den Wechsel von der Anwesenheit eines negativen Reizes hin zu seiner Abwesenheit (= negative Belohnung/Verstärkung) oder durch den Wechsel von der Abwesenheit eines positiven Reizes hin zu seiner Anwesenheit (= positive Belohnung/Verstärkung). Eine Bestrafung ist die Folge des Wechsels entweder von der Anwesenheit eines positiven Reizes hin zu dessen Abwesenheit (= negative Bestrafung) oder von der Abwesenheit eines negativen Reizes hin zu seiner Anwesenheit (= positive Bestrafung). Die Attribuierung einer Belohnung oder Bestrafung als negativ oder positiv ist dabei unabhängig von der Bewertung des Reizes als positiv oder negativ, sie richtet sich allein nach der Art des Reizwechsels von der Abwesenheit hin zu dessen Anwesenheit (= positiv) bzw. dessen Anwesenheit hin zur Abwesenheit (= negativ). Belohnungen führen zu einem verstärkten Auftreten des belohnten Verhaltens, Bestrafungen zur Abschwächung des Auftretens des bestraften Verhaltens.

2 Differenzfunktion

Wir modellieren Lernen, indem wir Ideen von B. F. Skinner aufgreifen und unter Zuhilfenahme des Konzepts der *Verhaltens-Bedeutungs-Paare* mathematisch formalisieren. Nach Skinner lässt sich Verhalten (Behavior, B) als eine Funktion von Vorbedingungen (Antecedents, A) und

(beabsichtigten oder unbeabsichtigten) Wirkungen (Consequences, C) auffassen. Wir kommen so zu einem *ABC*-Schema der Form

$$A \quad \longrightarrow \quad B \quad \longrightarrow \quad C.$$

Den Bestandteilen des ABC-Schemas ordnen wir *semantische Anker* zu, die sich von Zeitpunkt zu Zeitpunkt ändern können. Ein semantischer Anker ist, per definitionem, eine subjektive Referenz auf eine subjektive Wirklichkeit.

In unserem Fall ist das Subjekt der Lernende. In der basalen Variante (d.h. nur mit einem Reiz und ohne zusätzliche Gewichtung der Bewertung) besteht die subjektive Wirklichkeit des Lernenden aus der positiven $(+1)$ oder negativen (-1) Bewertung eines (durch einen Lehrenden oder die Umwelt hervorgerufenen) Reizes. Ist der Reiz nicht vorhanden, so weisen wir ihm den Wert 0 zu. Im mathematischen Modell verwenden wir eine Bewertungsfunktion β, die jedem Reiz r und jeder Situation s (vor oder nach einem Verhalten) eine Bewertung $\in \{+1,-1,0\}$ zuordnet:

$$\beta(r,s) := \begin{cases} 0 & \text{falls der Reiz } r \text{ in der Situation } s \text{ nicht vorhanden ist,} \\ +1 & \text{falls der Reiz } r \text{ in der Situation } s \text{ positiv empfunden wird,} \\ -1 & \text{falls der Reiz } r \text{ in der Situation } s \text{ negativ empfunden wird.} \end{cases}$$

Als semantische Anker betrachten wir zu gegebenem Reiz r und gegebener Situation s drei subjektive Referenzen: die Tripel $(r,s,0)$, $(r,s,+1)$ und $(r,s,-1)$. Da wir drei semantische Anker in Betracht ziehen, dies jedoch sowohl für die Antecedents A als auch die Consequences C tun, gibt es zu jedem Verhalten B und jedem Reiz r neun verschiedene Bedeutungen, modelliert als geordnete Paare semantischer Anker bestehend aus der Bewertung von r vor (Antecedens A) und nach dem (Consequences C) betrachteten Verhalten:

$$(\beta(r,A),\beta(r,C)) \quad \in \quad \{(+1,+1),(+1,-1),(+1,0),\ldots\}.$$

Für das Lernen entscheidend ist die Differenz aus Consequences C und Antecedents A, die sich formal als über den Reiz r parametrisierte Funktion des Verhaltens b auffassen lässt:

$$\Delta_r(B) = \beta(r,C) - \beta(r,A).$$

Ein *Verhaltens-Bedeutungs-Paar* besteht in der basalen Variante mit einem festen Reiz r aus der Deskription eines konkreten Verhaltens B des Lernenden und der sich daraus ergebenden Differenz:

$$\Big(B, \Delta_r(B)\Big).$$

Ist $\Delta_r(B) > 0$, so wird das Verhalten vom Lernenden als positiv erfahren, sodass es häufiger auftreten wird, ist $\Delta_r(B) < 0$, so wird das Verhalten vom Lernenden als negativ erfahren und seltener auftreten. Für das Lernen ist nur das Vorzeichen des Funktionswerts $\Delta_r(B)$ relevant. Dadurch bleiben von den neun möglichen Bedeutungen nur noch drei übrig:

$$\Delta_r(B) > 0, \quad \Delta_r(B) = 0 \quad \text{und} \quad \Delta_r(B) < 0.$$

Ein Lernvorgang besteht aus der Entwicklung eines Verhaltens-Bedeutungs-Paars. Wir assoziieren nun zu jedem Zeitpunkt k ein ABC-Schema

$$A_k \quad \longrightarrow \quad B_k \quad \longrightarrow \quad C_k.$$

Fokussiert auf einen Reiz r führt das zu einer reizspezifischen Differenz

$$\Delta_r(B_k) = \beta(r,C_k) - \beta(r,A_k)$$

und zum Verhaltens-Bedeutungs-Paar $(B_k, \Delta_r(B_k))$.

3 Ein beispielhaftes Szenario

Zu Demonstrationszwecken seien mehrere Fälle angenommen. Als Szenario soll ein Lernender, hier (der minderjährige) Max, dienen, der sein Zimmer nicht (regelmäßig) aufräumt und ein (nicht weiter definierter, erziehungsberechtigter) Lehrender, dessen Lehrziel es ist, dass Max sein Zimmer regelmäßiger aufräumt. Die vom Lehrenden zur Belohnung (Verstärkung) oder Bestrafung eingesetzten Reizänderungen betreffen einerseits die Gewährung von Taschengeld (G) bzw. dessen Entzug ($\neg G$), andererseits die Verhängung von Stubenarrest (S) bzw. dessen Absenz ($\neg S$). Ferner nehmen wir an, dass der Erhalt (also die Anwesenheit) von Taschengeld G in jeder Situation s positiv bewertet wird,

$$\beta(G,s) = +1,$$

die Verhängung (Anwesenheit) von Stubenarrest S in jeder Situation s hingegen negativ bewertet wird,

$$\beta(S,s) = -1.$$

Fall 1 (= negative Bestrafung)

Wir nehmen an, dass der Lernende Max sein Taschengeld erhält ($G \in A$), aber sein Zimmer nicht aufräumt ($\neg Z \in B$), und dass der Lehrende ihm daraufhin das Taschengeld entzieht ($\neg G \in C$). Max verliert also einen positiv bewerteten Reiz, was zu einer negativen Differenz führt:

$$\Delta_G(\neg Z) = \beta(\neg G,C) - \beta(G,A) = -1 - 0 = -1.$$

Der Lerneffekt ist, dass Max sein Zimmer seltener nicht aufräumt.

Fall 2 (= positive Bestrafung)

Wir nehmen an, dass der Lernende Max sein Zimmer nicht aufräumt ($\neg Z \in B$), und dass der Lehrende daraufhin Stubenarrest verhängt ($S \in C$, aber $\neg S \in A$). Max erhält also einen negativ bewerteten Reiz, was zu einer negativen Differenz führt:

$$\Delta_S(\neg Z) = \beta(S,C) - \beta(\neg S,A) = -1 - 0 = -1.$$

Der Lerneffekt ist, dass Max sein Zimmer seltener nicht aufräumt.

Fall 3 (= negative Belohnung)

Wir nehmen an, dass der Lernende Max Stubenarrest hat ($S \in A$), dabei sein Zimmer aufräumt ($Z \in B$), und dass der Lehrende ihm daraufhin den Stubenarrest erlässt ($\neg S \in C$). Max verliert also einen negativ bewerteten Reiz, was zu einer positiven Differenz führt:

$$\Delta_S(Z) = \beta(\neg S,C) - \beta(S,A) = 0 - (-1) = +1.$$

Der Lerneffekt ist, dass Max sein Zimmer häufiger aufräumt.

Fall 4 (= positive Belohnung)

Wir nehmen an, dass der Lernende Max kein Taschengeld erhält ($\neg G \in A$), sein Zimmer aufräumt ($Z \in B$), und dass der Lehrende ihm daraufhin Taschengeld gibt ($G \in C$). Max erhält also einen positiv bewerteten Reiz, was zu einer positiven Differenz führt:

$$\Delta_G(Z) = \beta(G,C) - \beta(\neg G,A) = +1 - 0 = +1.$$

Der Lerneffekt ist, dass Max sein Zimmer häufiger aufräumt.

Fall 5 (= zwei Reize und Summenbildung)

Wir betrachten nun das Lernen des Verhaltens „Zimmer aufräumen“ an drei aufeinander folgenden Zeitpunkten $k = 0$, $k = 1$ und $k = 2$ mit zwei unterschiedlichen Reizen, nämlich Stubenarrest erhalten (S) oder nicht erhalten ($\neg S$) und Taschengeld erhalten (G) oder nicht erhalten ($\neg G$).

Der Zeitpunkt $k = 0$ entspricht dem zweiten Fall, also dem Verhängen von Stubenarrest, weil Max sein Zimmer nicht aufgeräumt hat. Antecedents, Behavior und Consequences erhalten die Mengen semantischer Anker

$$A_0 := \{\neg S, \neg G\}, \quad B_0 := \{\neg Z\} \quad \text{und} \quad C_0 := \{S, \neg G\}$$

sowie die vier Reizbewertungen

$$\beta(\neg S, A_0) = 0, \quad \beta(\neg G, A_0) = 0, \quad \beta(S, C_0) = -1 \quad \text{und} \quad \beta(\neg G, C_0) = 0.$$

Daraus ergeben sich die beiden Differenzen

$$\begin{aligned} \Delta_S(\neg Z) &= \beta(S, C_0) - \beta(\neg S, A_0) = -1 - 0 = -1 \quad \text{und} \\ \Delta_G(\neg Z) &= \beta(\neg G, C_0) - \beta(\neg G, A_0) = 0 - 0 = 0 \end{aligned}$$

sowie deren Summe

$$\Delta(\neg Z) = \Delta_S(\neg Z) + \Delta_G(\neg Z) = -1 + 0 = -1.$$

Wegen $\Delta(\neg Z) < 0$ wird das Verhalten $\neg Z$ im Lernschritt zum Zeitpunkt $k = 0$ also abgeschwächt.

Der Zeitpunkt $k = 1$ entspricht dem dritten Fall, also dem Wegfall des Stubenarrests, da Max sein Zimmer aufgeräumt hat. Wir erhalten für Antecedents, Behavior und Consequences die Mengen semantischer Anker

$$A_1 := C_0 = \{S, \neg G\}, \quad B_1 := \{Z\} \quad \text{und} \quad C_1 := \{\neg S, \neg G\}$$

sowie die vier Reizbewertungen

$$\beta(S, A_1) = -1, \quad \beta(\neg G, A_1) = 0, \quad \beta(\neg S, C_1) = 0 \quad \text{und} \quad \beta(\neg G, C_1) = 0.$$

Daraus ergeben sich die beiden Differenzen

$$\Delta_S(Z) = \beta(\neg S, C_1) - \beta(S, A_1) = +1 \quad \text{und} \quad \Delta_G(Z) = \beta(\neg G, C_1) - \beta(\neg G, A_1) = 0$$

sowie deren Summe $\Delta(Z) = +1 > 0$. Das Verhalten Z wird also verstärkt.

Der Zeitpunkt $k = 2$ entspricht dem vierten Fall, also dem Erhalt von Taschengeld nach dem Aufräumen des Zimmers. Wir erhalten

$$A_2 := C_1 = \{\neg S, \neg G\}, \quad B_2 := \{Z\} \quad \text{und} \quad C_2 := \{\neg S, G\}$$

sowie die vier Reizbewertungen

$$\beta(\neg S, A_2) = 0, \quad \beta(\neg G, A_2) = 0, \quad \beta(\neg S, C_2) = 0 \quad \text{und} \quad \beta(G, C_2) = +1.$$

Daraus ergeben sich die beiden Differenzen $\Delta_S(Z) = 0$ und $\Delta_G(Z) = +1$ sowie deren Summe $\Delta(Z) = +1 > 0$. Das Verhalten Z wird also weiter verstärkt.

Fall 6 (= gleichzeitige Änderung zweier Reize und Weder-Belohnung-noch-Bestrafung)

Zum Zeitpunkt $k = 0$ betrachten wir eine Kombination aus positiver und negativer Bestrafung, nämlich

$$A_0 := \{\neg S, G\}, \quad B_0 := \{\neg Z\} \quad \text{und} \quad C_0 := \{S, \neg G\}.$$

Daraus resultieren die Reizbewertungen

$$\beta(\neg S, A_0) = 0, \quad \beta(G, A_0) = +1, \quad \beta(S, C_0) = -1 \quad \text{und} \quad \beta(\neg G, C_0) = 0$$

und die Differenzen $\Delta_S(\neg Z) = \Delta_G(\neg Z) = -1$ mit der Summe $\Delta(\neg Z) = -2 < 0$. Damit wird das Verhalten $\neg Z$ abgeschwächt.

Zum Zeitpunkt $k = 1$ betrachten wir den Fall, dass Max das richtige Verhalten zeigt, was jedoch seitens des Erziehers keine Änderung seiner Maßnahmen (Stubenarrest und kein Taschengeld) mit sich bringt, in unserer Notation also

$$A_1 := C_0 = \{S, \neg G\}, \quad B_1 := \{Z\} \quad \text{und} \quad C_1 := \{S, \neg G\}.$$

Daraus resultieren die Reizbewertungen

$$\beta(S, A_1) = -1, \quad \beta(\neg G, A_1) = 0, \quad \beta(S, C_1) = -1 \quad \text{und} \quad \beta(\neg G, C_1) = 0$$

und die Differenzen $\Delta_S(Z) = \Delta_G(Z) = 0$ mit der Summe $\Delta(Z) = 0$. Aufgrund des Differenzoperators ist dies (auf den ersten Blick kontraintuitiv) nicht gleichbedeutend mit einer Bestrafung, da C_1 gleich A_1 ist. Wenn wir aber annehmen, dass bei Summe 0 keine Änderung des Verhaltens eintritt, macht der Erzieher alles richtig, da Z von ihm gewünscht wird.

Fall 7 (= Weder-Belohnung-noch-Bestrafung mit kontingenten Folgen)

Vor dem Hintergrund von Fall 6 stellt sich die Frage, ob ein intuitiver Lernvorgang auch dann möglich ist, wenn es bei Summe 0 zu einer Änderung des Verhaltens kommt. Daher soll nun (nach einer positiven Bestrafung zum Zeitpunkt $k = 0$),

$$\begin{aligned} &A_0 := \{\neg S, G\}, \quad B_0 := \{\neg Z\} \quad \text{und} \quad C_0 := \{S, G\}. \\ &\beta(\neg S, A_0) = 0, \quad \beta(G, A_0) = +1, \quad \beta(S, C_0) = -1 \quad \text{und} \quad \beta(G, C_0) = +1, \\ &\Delta_S(\neg Z) = -1 \quad \text{und} \quad \Delta_G(\neg Z) = 0, \quad \Delta(\neg Z) = -1 \end{aligned}$$

und einer dementsprechenden Abschwächung des Verhaltens $\neg Z$ (also einer Änderung hin zu Z), zum Zeitpunkt $k = 1$ eine Weder-Belohnung-noch-Bestrafung angenommen werden,

$$\begin{aligned} &A_1 := \{S, G\}, \quad B_1 := \{Z\} \quad \text{und} \quad C_1 := \{S, G\}. \\ &\beta(S, A_1) = -1, \quad \beta(G, A_1) = +1, \quad \beta(S, C_1) = -1 \quad \text{und} \quad \beta(G, C_1) = +1, \end{aligned}$$

die aufgrund der Differenzen $\Delta_S(Z) = 0$ und $\Delta_G(Z) = 0$ mit der Summe $\Delta(Z) = 0$ nun im Gegensatz zu Fall 6 zu einer Änderung des Lernendenverhaltens zum Zeitpunkt $k = 2$ führt. Dem nun auftretenden nicht-gewünschten Verhalten $\neg Z$ kann der Lehrende in Form einer negativen Bestrafung begegnen, indem er Max das Taschengeld streicht,

$$\begin{aligned} &A_2 := \{S, G\}, \quad B_2 := \{\neg Z\} \quad \text{und} \quad C_2 := \{S, \neg G\}. \\ &\beta(S, A_2) = -1, \quad \beta(G, A_2) = +1, \quad \beta(S, C_2) = -1 \quad \text{und} \quad \beta(G, C_2) = 0, \end{aligned}$$

was aufgrund der Differenzen $\Delta_S(\neg Z) = 0$ und $\Delta_G(\neg Z) = -1$ mit der Summe $\Delta(\neg Z) = -1 < 0$ zur gewünschten Abschwächung von $\neg Z$ führt. Ein Änderung des Verhaltens trotz Summe 0

(im Falle einer Weder-Belohnung-noch-Bestrafung) zeigt damit keine kontraintuitiven Folgen, da durch Eingriff des Lehrenden (hier einer negativen Bestrafung), das Verhalten des Lernenden wieder dem vom Lehrenden intendierten Verhalten angepasst werden kann. Ein kontingentes Verhalten des Lernenden bei Summe 0 ist demnach im Modell konsistent abbildbar. Eine (kontraintuitive) Annahme einer Nicht-Änderung des Lernendenverhaltens im Falle der Summe 0 ist daher nicht notwendig.

Fall 8 (= gleichzeitige doppelte Reizänderung mit widersprüchlichem Lehren)

Auf die Summe 0 kommt man nicht nur, wenn ein Lernendenverhalten weder belohnt noch bestraft wird. Es ist auch im Falle einer (ob intendierten oder nicht-intendierten) widersprüchlichen Reizänderung möglich. Zum Zeitpunkt $k = 0$ betrachten wir daher eine Kombination aus unbeabsichtigter positiver Belohnung, die eigentlich als positive Bestrafung intendiert war, und einer negativen Bestrafung:

$$A_0 := \{\neg S, G\}, \quad B_0 := \{\neg Z\} \quad \text{und} \quad C_0 := \{S, \neg G\}.$$

Da Max nun als Nerd angenommen wird, bewertet er den Stubenarrest positiv (da er sowieso am liebsten in der Stube am PC hockt). Den Entzug des Taschengeldes bewertet er hingegen erwartungsgemäß negativ. Es gilt

$$\beta(\neg S, A_0) = 0, \quad \beta(G, A_0) = +1, \quad \beta(S, C_0) = +1 \quad \text{und} \quad \beta(\neg G, C_0) = +1,$$

was zu den Differenzen $\Delta_S(\neg Z) = +1$ und $\Delta_G(\neg Z) = -1$ mit der Summe $\Delta(Z) = 0$ führt. Da bei der Beurteilung des Falles 7 ein kontingentes Lernendenverhalten bei Summe 0 angenommen wurde, muss der Fall 8 hinsichtlich eines Falles 8.1 und 8.2 unterschieden werden – einmal ohne Änderung des Lernendenverhaltens, einmal mit Änderung des Lernendenverhaltens.

Fall 8.1 (= Fall 8 ohne Änderung des Lernendenverhaltens)

Der Lehrende merkt anhand der Nicht-Änderung des Lernendenverhaltens, dass die Änderung der Reize keine Wirkung gezeigt hat und schließt darauf, dass Nerds Stubenarrest postiv bewerten. Dementsprechend erlässt er Max zum Zeitpunkt $k = 1$ den Stubenarrest, setzt also mit einer positiven Bestrafung an:

$$A_1 := C_0 = \{S, \neg G\}, \quad B_1 := \{\neg Z\} \quad \text{und} \quad C_1 := \{\neg S, \neg G\}$$

mit den Reizbewertungen

$$\beta(S, A_1) = +1, \quad \beta(\neg G, A_1) = 0, \quad \beta(\neg S, C_1) = 0 \quad \text{und} \quad \beta(\neg G, C_1) = 0$$

und den Differenzen $\Delta_S(\neg Z) = -1$ und $\Delta_G(\neg Z) = 0$ mit der Summe $\Delta(Z) = -1 < 0$. Das Verhalten $\neg Z$ wird entsprechend abgeschwächt, was auch der Intention des Lehrenden entspricht.

Fall 8.2 (= Fall 8 mit Änderung des Lernendenverhaltens)

Der Lehrende denkt aufgrund der Änderung des Lernendenverhaltens irrtümlich, dass die Änderung der Reize seine Wirkung gezeigt hat. Entsprechend kann er eine weitere Reizänderung unterlassen (=8.2.1):

$$A_1 := C_0 = \{S, \neg G\}, \quad B_1 := \{Z\} \quad \text{und} \quad C_1 := \{S, \neg G\}$$

mit den Reizbewertungen

$$\beta(S,A_1) = +1, \quad \beta(\neg G,A_1) = 0, \quad \beta(S,C_1) = +1 \quad \text{und} \quad \beta(\neg G,C_1) = 0$$

und den Differenzen $\Delta_S(Z) = 0$ und $\Delta_G(Z) = 0$ mit der Summe $\Delta(Z) = 0$. Da der Lernende weder zum Zeitpunkt $k = 1$ noch zum Zeitpunkt $k = 0$ etwas gelernt hat, ist sein Verhalten nicht stabil, was sich auch durch die im Modell angenommene Kontingenz des Lernendenverhaltens widerspiegelt. Bei Auftreten des nicht-erwünschten Verhaltens (z.B. zum Zeitpunkt $k = 2$) wird der Lehrende eine weitere (positive oder negative) Bestrafung vornehmen müssen. Dies wird er jedoch mittels eines dritten Reizes unternehmen müssen, da der Lehrende aufgrund der erstmaligen Verhaltensänderung des Lernenden annehmen muss, dass sein ursprüngliches Lehrverhalten effektiv war. Das gilt allerdings nur für den Fall, dass sich der Lehrende zum Zeitpunkt $k = 1$ dafür entschließt den Lernenden weder zu bestrafen noch zu belohnen.

Hätte sich der Lehrende hingegen zum Zeitpunkt $k = 1$ entschieden, das Verhalten des Lernenden zu belohnen (=8.2.2), so kann dies, abhängig von der Reizänderung unterschiedliche Folgen haben. Sollte sich der Lehrende für eine positive Belohnung des Lernenden durch Gewährung von Taschengeld entscheiden (=8.2.2.1),

$$A_1 := C_0 = \{S, \neg G\}, \quad B_1 := \{Z\} \quad \text{und} \quad C_1 := \{S, G\}$$

mit den Reizbewertungen

$$\beta(S,A_1) = +1, \quad \beta(\neg G,A_1) = 0, \quad \beta(S,C_1) = +1 \quad \text{und} \quad \beta(G,C_1) = +1,$$

so würde das zu den Differenzen $\Delta_S(Z) = 0$ und $\Delta_G(Z) = +1$ mit der Summe $\Delta(Z) = +1 > 0$ führen und damit tatsächlich zur Verstärkung des gewünschten Lernendenverhaltens.

Der Lehrende hätte sich aber auch für eine vermeintliche negative Belohnung entscheiden können, indem er Max den Stubenarrest erlässt (=8.2.2.2):

$$A_1 := C_0 = \{S, \neg G\}, \quad B_1 := \{Z\} \quad \text{und} \quad C_1 := \{\neg S, \neg G\}.$$

Da der Lernende dies als negative Bestrafung ansehen würde,

$$\beta(S,A_1) = +1, \quad \beta(\neg G,A_1) = 0, \quad \beta(\neg S,C_1) = 0 \quad \text{und} \quad \beta(\neg G,C_1) = 0,$$

und damit auch den Differenzen $\Delta_S(Z) = -1$ und $\Delta_G(Z) = 0$ mit der Summe $\Delta(Z) = -1 < 0$, würde der Lernende sein Verhalten ändern, sodass das nicht gewünschte Verhalten auftritt und der Lehrende analog zum Fall 8.1 darauf schließen müsste, dass Nerds Stubenarrest positiv bewerten.

Abschließendes Fazit: Es ist zu sehen, dass das Modell auch komplexere Lernszenarien adäquat, da intuitiv und funktional, abzubilden vermag.

4 Stevenssches Potenzgesetz

Wir sind bisher von den gleichen Grundannahmen wie Mnih und Kollegen in [2, 3] ausgegangen, namentlich dass (1) sich die Belohnung aus einer *Differenz* der Bewertungen auf einander folgender und durch den Agenten (im Beispiel: Max) beeinflusster Situationen ergibt und (2) die Belohnungen dreiwertig – also -1 für negativ, 0 für neutral und $+1$ für positiv – sind.

An dieser Stelle erscheint die Verwendung von Bewertungs*differenzen* nach dem Stevensschem Potenzgesetz in seiner differenziellen Form [4]

$$\frac{\Delta E}{E} = k \cdot \frac{\Delta R}{R},$$

wobei R für eine Reizgröße, E für eine Empfindungsgröße und k für eine reellwertige Konstante stehen, reizpsychologisch plausibel. In unserem Szenario bestünde der Reiz aus der Gewährung oder Verweigerung von Taschengeld bzw. aus dem Verhängen oder Nicht-Verhängen von Stubenarrest. Die sich daraus ableitende Empfindung wäre die Bewertung der Situation. Wir erhalten für unseren Fall

$$\frac{E(C)-E(A)}{E(A)} = k \cdot \frac{R(C)-R(A)}{R(A)},$$

wobei $R(A)$ und $E(A)$ für Reiz- bzw. Empfindungs*größen* bei Vorliegen der Antecedents und $R(C)$, $E(C)$ Reiz- bzw. Empfindungsgrößen bei Vorliegen der Consequences bedeuten. Für die Empfindungsdifferenz ergibt sich:

$$\Delta E = E(C)-E(A) = k \cdot \frac{R(C)-R(A)}{R(A)} \cdot E(A).$$

Um in unseren Beispielen Nullen im Nenner zu vermeiden, müssen wir die tatsächlichen Reiz- und Empfindungsgrößen mittels einer Exponentialfunktion aus unseren Reizbewertungen errechnen. Bezüglich des Taschengeldes setzen wir $E(\neg G) = 2^0 = 1$, $E(G) = 2^1 = 2$, $R(\neg G) = 2^0 = 1$, $R(G) = 2^1 = 2$ sowie $k = 1$ und erhalten

A	C	$E(A)$	$R(A)$	$R(C)$	ΔE
$\neg G$	$\neg G$	1	1	1	0
$\neg G$	G	1	1	2	1
G	$\neg G$	2	2	1	-1
G	G	2	2	2	0

Bezüglich des Stubenarrestes setzen wir $E(\neg S) = 2^0 = 1$, $E(S) = 2^{-1} = 1/2$, $R(\neg S) = 2^0 = 1$, $R(S) = 2^{-1} = 1/2$ sowie $k = 2$ und erhalten

A	C	$E(A)$	$R(A)$	$R(C)$	ΔE
$\neg S$	$\neg S$	$1/2$	1	1	0
$\neg S$	S	1	1	$1/2$	-1
S	$\neg S$	$1/2$	$1/2$	1	1
S	S	$1/2$	$1/2$	$1/2$	0

5 Ausblick: Quantenlogik

Die mathematischen Modelle der Quantenmechanik ermöglichen über die Quantenlogik [5, 6] eine erstaunliche Verbindung zum hier betrachteten Lernen durch Bewertungsdifferenzen von Reizgrößen. Dies soll nun kurz skizziert werden.

Zur Konstruktion des mathematischen Modells nehmen wir die beiden Verhaltensmöglichkeiten Z und $\neg Z$ als Ket-Vektoren $|Z\rangle$ und $|\neg Z\rangle$ und betrachten diese als Orthonormalbasis eines Euklidischen Raums. Der Lernfortschritt wird jeweils durch einen Vektor $|L\rangle$ in diesem Raum kodiert. Beginnend mit dem Nullvektor $|L_0\rangle = 0$ addieren wir bei jedem Lernschritt im Fall des Verhaltens Z den Vektor $\Delta(Z) \cdot |Z\rangle$ und im Fall der Verhaltens $\neg Z$ den Vektor $\Delta(\neg Z) \cdot |\neg Z\rangle$. Nach dem k-ten Lernschritt erhalten wir einen Vektor $|L_k\rangle$. Um herauszufinden, was gelernt wurde, vergleichen wir die beiden Skalarprodukte

$$\langle Z|L_k\rangle \quad \text{und} \quad \langle \neg Z|L_k\rangle.$$

Im Fall $\langle Z|L_k\rangle > \langle \neg Z|L_k\rangle$ ist das Verhalten Z gelernt, im Fall $\langle \neg Z|L_k\rangle > \langle Z|L_k\rangle$ wurde dagegen das Verhalten $\neg Z$ gelernt.

Literatur

[1] GERRIG, R. J.: *Psychologie*. Pearson, 2014. Kapitel 6.3 Operantes Konditionieren: Lernen von Konsequenzen, S. 216–228.

[2] MNIH, V., K. KAVUKCUOGLU, D. SILVER, A. GRAVES, I. ANTONOGLOU, D. WIERSTRA, und M. RIEDMILLER: *Playing atari with deep reinforcement learning*. In *NIPS Deep Learning Workshop*. 2013.

[3] MNIH, V., K. KAVUKCUOGLU, D. SILVER, A. A. RUSU, J. VENESS, M. G. BELLEMARE, A. GRAVES, M. RIEDMILLER, A. K. FIDJELAND, G. OSTROVSKI, S. PETERSEN, C. BEATTIE, A. SADIK, I. ANTONOGLOU, H. KING, D. KUMARAN, D. WIERSTRA, S. LEGG, und D. HASSABIS: *Human-level control through deep reinforcement learning*. *Nature*, 518, S. 529–533, 2015.

[4] STEVENS, S. S.: *On the psychophysical law*. *Psychol Rev*, 64(3), S. 153–181, 1957.

[5] BIRKHOFF, G. und J. VON NEUMANN: *The logic of quantum mechanics*. *Annals of Mathematics*, 2nd Ser. 37(4), S. 823–843, 1936.

[6] WIRSCHING, G., M. WOLFF, und I. SCHMITT: *Quantenlogik - Eine Einführung für Ingenieure und Informatiker*. Springer, 2020 (in Vorbereitung).