DOI: 10.5771/0342-300X-2020-3-193

Digitale Beschäftigtenratings in der tertiären Arbeitswelt

Insbesondere in Unternehmen des Dienstleistungssektors halten neue Instrumente des „algorithmischen Managements“ vermehrt Einzug, welche auf die softwaregestützte Bewertung von Beschäftigten durch andere Beschäftigte abzielen. Das Management propagiert diese digitalen, multiperspektivischen Bewertungssysteme als moderne und verlässliche Alternative zur Beurteilung durch direkte Vorgesetzte, aus der gerechtere Entlohnung und fairere Verteilung von Aufstiegschancen resultieren. Doch anhand der Analyse eines besonders umfassenden und avancierten Systems bei einem großen Versandhändler lässt sich zeigen, dass das Beschäftigtenrating intransparent ist, zu mehr Leistungsdruck führt und das Risiko für neue Ungleichheiten im betrieblichen Kontext steigert.

PHILIPP STAAB, SASCHA-CHRISTOPHER GESCHKE

1 Einleitung

Digitale Ratings sind ein zentrales Element der Herstellung von Konsumentenvertrauen im kommerziellen Internet. Insbesondere E-Commerce-Plattformen sind davon abhängig, dass Anbieter*innen und Konsument*innen in einen freiwilligen Austausch miteinander treten. Wie auf allen Märkten ist auch auf digitalen Marktplätzen soziales Vertrauen hierfür eine Voraussetzung, ohne die kein Kunde und keine Kundin Geld an einen Händler überweisen würde. Durch die unpersönlichen, „delokalisierten“ (Kirchner/Beyer 2016) Interaktionsstrukturen des kommerziellen Internets ist solches Vertrauen allerdings zunächst einmal nicht gegeben: Wer das Produkt eines Dritten auf Ebay, Amazon oder Aliexpress kauft, weiß nicht, ob er es nicht mit Betrügern zu tun hat, die sich mit dem Kaufpreis aus dem Staub machen, ohne die gewünschte Ware zu versenden. Zur Lösung dieses Problems wurde im kommerziellen Internet eine spezifische, Disziplin erzeugende und damit Vertrauen verbürgende Praxis etabliert: horizontale Ratings. Über Konsumentenbewertungen *(consumer ratings)* erarbeiten sich Anbieter*innen ihre Seriosität.

Als Bestandteil von Strategien des „algorithmischen Managements“ (vgl. Beverungen 2017; Staab 2019, S. 233ff.) in der Arbeitswelt werden digitale Ratings nun zunehmend auch als betriebliche Evaluierungs- und Kontrollinstrumente im Kontext unterschiedlicher tertiärer Tätigkeiten eingesetzt. Technologien des algorithmischen Managements kommen heute in so unterschiedlichen Dienstleistungsbereichen wie dem Einzelhandel (Evans/Kitchin 2018), der Kreativ- und Büroarbeit (Moore 2018a, 2018b; Moore/Piwek 2017) oder der sogenannten Gig-Economy (Lee/Baykal 2017; Lee 2015; Schor/Attwood-Charles 2017; Schor et al. 2017; Attwood-Charles 2019) zum Einsatz. Ratingsysteme sind dabei neben der Strukturierung von Interfaces und konstanter Überwachung von Tätigkeiten (Tracking) eine der drei Kernkomponenten (Staab 2019, S. 233ff.). Bei ihrer Implementierung in der Arbeitswelt sind sie anschlussfähig an etabliertere Strategien der Leistungsbeurteilung, wie etwa das 360-Grad-Feedback[1] (vgl. Bröckling 2003, S. 83ff.).

Wenig überraschend sind Internet- und insbesondere E-Commerce-Unternehmen die Vorreiter beim Einsatz neuer Digitalratings. So sorgte beispielsweise 2015 ein Bericht der „New York Times“ für Aufsehen, in dem u. a. das *Anytime-Feedbacktool* beschrieben wurde, das damals im

1 Das 360-Grad-Feedback ist ein Instrument aus dem Bereich des Human Resource Managements, das ursprünglich der Auswahl von Offiziersanwärtern in der Wehrmacht diente, und zudem ein Vorläufer des heutigen Assessment-Centers (vgl. Fleenor/Prince 1997, S. 51).

Amazon-Hauptquartier in Seattle zum Einsatz kam (Kantor/Streitfeld 2015). Die Software ermöglicht den Beschäftigten, ihre Kolleginnen und Kollegen zu evaluieren und diese Bewertungen an das Management zu übermitteln. Die über die Software gesammelten Informationen fließen direkt in Leistungsbeurteilungen ein und sind ein wichtiger Bestandteil der Legitimierung der *Hire-and-Fire*-Kultur, die das Unternehmen prägt (ebd.). Weniger bekannt ist, dass auch andere große Software- und Internetfirmen Rating-Technologien für Unternehmen im Angebot haben. Google bietet beispielsweise unter dem Stichwort *re:work Tools* eine eigene webbasierte Plattform an, auf der Materialien und Strategien zu Themenbereichen wie Zielsetzung des Unternehmens, Personalentscheidungen, *People Analytics* und Teambuilding kostenfrei zur Verfügung gestellt werden. In Deutschland haben sich zahlreiche Firmen an diesen Anleitungen bedient und setzen Anwendungen ein, die auf der Google-Philosophie basieren.

Um die spezifischen Funktionen und Effekte neuerer digitaler Rating-Instrumente für bzw. auf Arbeit zu erschließen, werden wir uns in diesem Beitrag einem besonders ambitionierten Projekt aus Deutschland widmen. Seit rund dreieinhalb Jahren kommt bei einem großen Internetversandhändler eine Anwendung zum Einsatz, die wir im Folgenden RADAR nennen werden. Hier werden – auf Basis umfassender, wechselseitiger Bewertungen unter Kolleg*innen, welche in die Software eingegeben und verarbeitet werden – Leistungsbeurteilungen (sogenannte Scores) für die einzelnen Beschäftigten erstellt. Wir werden RADAR zunächst nach bestem Wissen beschreiben (Abschnitt 2). Dabei beziehen wir uns auf Daten aus einer von uns durchgeführten und von der Hans-Böckler-Stiftung finanzierten Fallstudie, die zwischen 2017 und 2019, während der Implementierungsphase von RADAR, erhoben wurden. Im Kontext dieser explorativen Untersuchung haben wir Schulungsmaterialien zur betreffenden Software analysiert, zwei Pretest- und zehn Beschäftigteninterviews geführt sowie zwei Gruppendiskussionen mit Angestellten realisiert.[2]

Beim Sampling der Interviewpartner*innen wurde die zergliederte Firmenstruktur des betreffenden Unternehmens berücksichtigt, um mögliche Unterschiede in den Erfahrungen der Beschäftigten mit RADAR einzufangen. Das heißt, es wurden keine Befragten ausgewählt, die aus derselben Abteilung stammten oder die gleiche oder eine ähnliche Position in der Hierarchie besetzten. Zudem haben wir uns, dem Leitbild des theoretischen Samplings (Glaser/Strauss et al. 1968) folgend, bemüht, durch die Auswahl der Interviewees ein möglichst breites Spektrum an Einstellungen zu RADAR abzudecken. Die Ergebnisse aus den vorangegangenen Analysen wurden schließlich in Experteninterviews mit Jurist*innen und Gewerkschafter*innen nochmals überprüft bzw. um deren Perspektiven erweitert. Trotz mehrfacher Anfragen war es leider zu keinem Zeitpunkt möglich, Gesprächspartner auf der Managementseite der betreffenden Firma zu akquirieren, um ebenso die Unternehmerperspektive einfließen zu lassen.

In Abschnitt 3 beschreiben wir am Fallbeispiel, in welcher Form und mit welchen Effekten digitale Ratings zur Leistungsvermessung und Kontrolle (3.1) eingesetzt werden können und welche Auswirkungen auf die Arbeitsqualität (3.2) sowie die betriebliche Sozialstruktur (3.3) sich dabei beobachten lassen. Unsere Ausführungen sind als explorative Sonde zu verstehen, die an einem Beispiel bestimmte Möglichkeitshorizonte des Einsatzes digitaler Ratings in der Dienstleistungsarbeit erkundet. Wie auch bei anderen Technologien ist gewiss von einer grundsätzlichen Gestaltbarkeit (vgl. z. B. Hirsch-Kreinsen et al. 2015) des jeweiligen Zuschnitts und Einsatzes von Ratings auszugehen, weshalb die von uns beschriebenen Entwicklungen sich nicht zwangsläufig bei jedem System einstellen müssen. Wir schließen mit einem kontextualisierenden Fazit und einigen Erwägungen zur Rolle von Mitbestimmung und Datenschutz im Zusammenhang mit neueren Formen von Digitalratings (4).

2 Was ist RADAR?

RADAR ist eine Software, die als Back-Office-Anwendung zum Einsatz kommt. Insgesamt dient sie der Führung und Verwaltung von mehreren tausend Mitarbeiter*innen, die zur regelmäßigen Nutzung des Systems und zur aktiven Teilnahme angehalten werden.[3] Offiziell geht es bei RADAR um die möglichst umfassende Einschätzung der Fähigkeiten und die Vermessung der Leistung der Angestellten. Beschäftigte nehmen es jedoch primär als Arbeitskontrolle wahr.

Um eine möglichst umfassende Evaluation einzelner Mitarbeiter*innen zu ermöglichen, fließen Informationen aus zwei verschiedenen Prozeduren in das System ein: Einerseits sind hochfrequente Echtzeitbewertungen Teil des Systems, die Kolleg*innen zu jeder Zeit und an jedem Ort ad hoc abgeben können.[4] Andererseits gibt es umfangrei-

2 Zur Entwicklung und Qualitätsverbesserung unseres Erhebungsinstruments wurden vorab zwei Beschäftigteninterviews geführt (Pretest), die nur zur Konstruktion des Fragebogens dienten und nicht in die Auswertungen selbst eingeflossen sind.

3 Bei Nichtteilnahme können Beschäftigte nicht an den über RADAR verteilten Aufstiegsmöglichkeiten partizipieren.

4 Dieses Echtzeitrating wurde in der Pilotphase in einer mobilen, App-basierten Form zwar getestet, letztlich wurden nach Angaben der von uns interviewten Beschäftigten jedoch mit RADAR 1.0 zwei Formen von Echtzeitratings eingesetzt, die an frei zugänglichen, stationären Arbeitsplätzen genutzt werden konnten.

che periodische Mitarbeiterbeurteilungen, die große Ähnlichkeit mit den *re:Work*-Tools von Google aufweisen und im Grunde die Funktion regelmäßiger Leistungsbeurteilungen haben, die klassischerweise von Vorgesetzten ausgearbeitet werden. Die periodische Evaluation vollzieht sich über einen Zeitraum von etwa eineinhalb Monaten: Zunächst nominieren alle Mitarbeiter*innen *kurz* vor der Feedback-Runde Personen, mit denen sie häufig und eng im letzten Turnus zusammengearbeitet haben, und werden von diesen umfassend beurteilt. Anschließend schreiben die direkten Vorgesetzten eine Stellungnahme zu den Evaluationen und reichen ein Datenpaket mit den gesammelten Informationen und Beurteilungen an ein Fallprüfungskomitee weiter. Dieser Ausschuss entscheidet letztlich über die Leistungsklassifikation und über die Aufstiegsperspektiven von Beschäftigten.

Im Zentrum beider Datenerhebungsprozeduren steht vor allem die *wechselseitige Bewertung der Leistung von Beschäftigten vergleichbarer Hierarchiestufen*. Wir sprechen daher im Folgenden von horizontalen Worker-Coworker-Ratings, da vertikale Kontrolle (Vorgesetzte – Mitarbeiter) in dieser Konstruktion auf eine horizontale Ebene (Mitarbeiter – Mitarbeiter) verschoben wird (vgl. Schaupp/Staab 2018). Jede Bewertung besteht aus einem quantifizierenden Rating, das durch eine qualitative Rezension (Freitext) argumentativ begründet und an Beispielen belegt werden muss. Ein großes Spektrum an Bewertungsmöglichkeiten ergibt sich aus den acht bis zwölf vom Unternehmen vorgegebenen abteilungsspezifischen Themenbereichen.

Die generierten Bewertungsdaten aus den Ratings werden anschließend aggregiert und über einen Algorithmus in individuelle *Mitarbeiter-Scores* zusammengefasst. Ob die Ratings lediglich aufsummiert und gemittelt werden oder ob eine Gewichtung erfolgt, wissen die Beschäftigten nicht. Ebenfalls unbekannt sind die Kriterien, nach denen die unzähligen qualitativen Rezensionen durch Vorgesetzte verarbeitet werden. Ein Fallprüfungskomitee erstellt am Ende der RADAR-Prozedur, auf der Basis dieser personenbezogenen Beschäftigtenbewertungen (Datenpakete), eine Einteilung der Belegschaft in drei Gruppen: Low-, Good- und Top-Performer. Diese Einteilung wird genutzt, um individuelle Bewertungsgespräche zu strukturieren, betriebliche Aufstiegsoptionen zu verteilen und Lohnsteigerungen zu gewähren beziehungsweise zu versagen.

Wie vergleichbare Produkte aus dem Bereich moderner Unternehmenssoftware wird RADAR in der betriebsöffentlichen Präsentation des Managements als ein Instrument beschrieben, das

1. *objektive und transparente Personalentscheidungen* ermöglichen solle, da das Management nun auf der Basis der Meinung jener Personen entscheiden könne, die direkt in den Arbeitsprozess integriert sind;
2. Wertschätzung kommuniziere und die Motivation der Mitarbeiter*innen steigere, da ihnen neue *Mitbestimmungskanäle* eröffnet würden;
3. *berechenbare Karriereoptionen* im Unternehmen auf Basis kompetenter, datengesteuerter Leitungsbeurteilungen möglich mache.

Die Darstellung des RADAR-Systems als ein Instrument der objektiven und validen Leistungsvermessung wird dadurch bekräftigt, dass Entwicklungsmöglichkeiten im Unternehmen nicht mehr allein vom Urteil einzelner Vorgesetzter abhängig sein, sondern *vermeintlich* auf einer möglichst breiten Bewertung der Arbeitsleistung durch Kolleg*innen beruhen. Transparente Kompetenzmodelle und Anforderungen an Berufsrollen, wie sie in den Ratingkategorien festgelegt sind, tragen dazu bei, dass RADAR von einzelnen Beschäftigten zunächst als ein fairer und vertrauenswürdiger Beurteilungsprozess wahrgenommen wird.

3 Perspektiven horizontaler Digitalratings

Im Verlauf der Implementierung des RADAR-Systems hat sich bei der großen Mehrheit unserer Gesprächspartner*innen der anfängliche Vertrauensvorschuss recht schnell erschöpft. Nicht Objektivität und Transparenz, sondern eine fragwürdige Leistungsvermessung und allgegenwärtige Kontrolle stehen ihrer Ansicht nach im Vordergrund (dazu Abschnitt 3.1). Statt mit einem Klima der Wertschätzung wird RADAR von ihnen mit einer Verschlechterung der Arbeitsqualität in Verbindung gebracht (3.2). Zudem scheint das System, unseren Informationen zufolge, keine planbaren Karriereoptionen zu schaffen und zu vermehren, wie es vom Unternehmen stets bekundet wird, sondern es funktioniert als Werkzeug der Stratifizierung und Lohnkontrolle (3.3).

3.1 Leistungsvermessung und Kontrolle

Ratingsysteme wie RADAR sind in einem sehr grundsätzlichen Sinne Instrumente der Leistungsvermessung. Nicht nur zielen sie im Kern auf die Erhebung leistungsbezogener Daten der Beschäftigten ab, die zur Beurteilung der Arbeit aggregiert werden. Sie erweitern auch den Zugriff des Managements auf leistungsbezogene Informationen: Vorgesetzte haben jetzt nicht mehr nur Zugriff auf Arbeitsergebnisse oder automatisch erhobene Kennzahlen (etwa Verkäufe oder produzierte Stückzahlen einzelner Beschäftigter). Sie verfügen über ein ganz neues und umfassendes Arsenal an sehr *persönlichen* Beurteilungsdaten, die direkt aus dem Arbeitsprozess stammen. Gerade für Dienstleistungstätigkeiten, die bis heute vielfach von einem hohen Maß „verantwortlicher Autonomie“ (Friedman 1987) geprägt sind, erschließen digitale Rating-

systeme dem Management also neue Überwachungs- und Sanktionsmöglichkeiten.

Dies zeigt sich exemplarisch bei RADAR, das ein System ersetzt hat, in dem es bis dahin meist nur wenig Prozessüberwachung durch Vorarbeiter*innen gab, die ferner kaum Kapazitäten hatten, direkt in die jeweiligen Tätigkeitsabläufe der einzelnen Mitarbeiter*innen zu intervenieren. Diese bisher lückenhafte, personengebundene Kontrollpraxis wird durch RADAR erheblich erweitert. Beschäftigte beschreiben das System beispielsweise als „360-Grad-Überwachung" oder als „System der kompletten Kontrolle". Bei ihnen entsteht häufig das Gefühl, der Bewertung durch Kolleg*innen zu jeder Zeit ausgeliefert zu sein und entsprechend unter permanenter Beobachtung zu stehen. Dies wird als Quelle von Arbeitsdruck und Stress identifiziert. So gibt ein*e Mitarbeiter*in zu Protokoll:

> „RADAR ist ein Feedback Tool, das den Mitarbeitern verkauft wird, um ihre Stärken zu sehen, zu erkennen und dementsprechend zu fördern. Aber ich hab's eher als ganz umgekehrt wahrgenommen. Eigentlich wird der Mitarbeiter dadurch komplett kontrolliert und klein gehalten, und durch dieses System werden dem Mitarbeiter viel mehr Steine in den Weg gelegt, als dass es in irgendeiner Form der Entwicklung dient."

Die Möglichkeit konstanter Bewertungen erzeugt ein Klima der Verunsicherung unter den von uns interviewten Beschäftigten, wie eine weitere Person im Interview am eigenen Beispiel schildert:

> „[…] also ich kann nicht einfach mal einen schlechten Tag haben, […] vielleicht ist es manchmal nur eine Kleinigkeit und je nachdem, mit wem ich da zu tun habe, und es bleibt nicht aus, dass da natürlich die Persönlichkeiten von Personen eine Rolle spielen, aber ein paar Monate später schlägt sich das in einem Feedback nieder, eine Situation, an die ich mich überhaupt nicht mehr erinnere."

Es zeigt sich an diesem *exemplarischen* Zitat recht deutlich, dass die Überwachungspraxis, die mit RADAR implementiert wird, zugleich disziplinierende Effekte zeitigt. Ratingsysteme wie RADAR sind daher nur ungenügend beschrieben, wenn man sie als Instrumente der Leistungsvermessung auffasst. Es handelt sich um Technologien, die systematisch Druck zur verstärkten Selbstdisziplinierung erzeugen und Hybride klassischer „direkter" und „indirekter" Kontrollformate bilden (vgl. Marrs 2018). *Indirekt,* weil kein unmittelbarer Zugriff durch Führungspersonal auf die Arbeitssituation erfolgt – *direkt,* weil diese Aufgaben, was den Effekt der Disziplinierung angeht, von Kollegen und Kolleginnen teilsubstituiert werden.

Auch die vom Management angepriesene Steigerung der Objektivität von Leistungsbeurteilungen durch die sogenannte Erweiterung der Beteiligungsmöglichkeiten von Beschäftigten sehen die von uns interviewten Betroffenen meist kritisch. Zum einen geben sie zu bedenken, dass der eigentliche Kern ihrer Arbeit stark von der jeweiligen Persönlichkeitsstruktur abhänge, die aus normativen Gründen keiner objektiven Beurteilung unterzogen werden sollte, da dabei im Prinzip der Mensch im Ganzen und nicht nur seine reine Arbeitsleistung losgelöst bewertet werde:

> „Ich glaube, man [kann] messen […], wie verrichtet jemand seine Arbeit, wie hat er Prozesse verstanden, und kann er das alles gründlich, selbstständig und so weiter erledigen. Aber ich finde, ein ganz großer Aspekt in Teamarbeit ist, wie arbeite ich mit jemandem zusammen? Da geht es auch um Persönlichkeiten. […] Es ist schwierig, da es eigentlich nichts ist, wo man eine Bewertung abgeben sollte, meiner Meinung nach, also weil es da um persönliche Eigenschaften geht."

Zum anderen mangelt es an Vertrauen in die Befähigung von Kolleg*innen, unter den Bedingungen der Konkurrenz um betriebliche Aufstiegschancen neutrale Leistungsbeurteilungen zu verfassen:

> „Ich traue den meisten kein Feedback zu. Und das ist eben mein allergrößter Kritikpunkt, […] da geht es natürlich auch um Konkurrenz. Also es gibt eben nur eine begrenzte Möglichkeit an Aufstiegschancen."

Darüber hinaus kritisieren einige Mitarbeiter*innen die recht deutliche Beeinflussung der Erhebungsverfahren durch das Management. Der Spielraum der Beschäftigten für freie und objektive Beurteilungen wird vom Design der Anwendung stark eingeschränkt. So ist insbesondere die Umformung von persönlichen Bewertungen in vermeintlich objektive Kennwerte (Scores), die den Wert der Mitarbeiter*innen im betriebswirtschaftlichen Sinne wiedergeben sollen, weitestgehend intransparent. Die Beschäftigten wissen nicht, welche Gewichtungen implizit eine Rolle spielen, und haben deshalb keine reelle Handlungsmacht: Sie können weder die Wertigkeit der eingehenden Informationen einschätzen noch die daraus resultierenden Ergebnisse der Scores (hierzu gleich mehr).

Über die Struktur der Bewertungsbögen und die Kontrolle der Auswertungsgewichtungen sind dem System nicht nur spezifische normative Standards eingeschrieben, die den Handlungsspielraum beschränken. Vielmehr erschließt sich aus der methodischen Begutachtung des Erhebungsinstruments recht schnell, dass die Ergebnisse der Evaluation durch das Design systematisch beeinflusst werden. Die vom Unternehmen konzipierten Bewertungsskalen begünstigen gezielt die Auswahl bestimmter Ausprägungen (niedriger Bewertungen). Konkreter formuliert: Das Erhebungsinstrument ist so angelegt, dass es absehbare Ergebnisse produziert, die im Interesse des Managements liegen. Planmäßig verringerte Werte in den Ratings senken die Mitarbeiter-Scores und rechtfertigen mögliche Sanktionen oder die Stagnation auf einer Position bzw. einem Lohnniveau anhand von vermeintlich objektiven und fairen Kennwerten.

Die Kategorien der Bewertungsskala werden durch das Management, in strategischer Absicht, mit bestimm-

ten Bezeichnungen versehen; mittels Schulungen und Präsentationen wirkt es zudem verstärkend darauf hin, dass die Kategorien in der erwünschten Weise wahrgenommen werden. So übt das Management Einfluss auf die Ergebnisse des Ratingverfahrens aus, wie ein*e Beschäftigte*r *exemplarisch* zu Protokoll gibt:

> „Ich weiß nur noch, dass es irgendwie absurd war, also die höchste Stufe ist halt: ‚übertrifft irgendwie sowohl innerhalb der Firma als auch außerhalb alles und ist einzigartig'. Also die höchste Stufe kann man eigentlich nicht erreichen. Also das ist schon so formuliert, dass es eigentlich Blödsinn ist und auch alle, also die zwei bis drei Stufen, die da drunter kommen, klingen immer noch so, als wenn derjenige fantastisch in dem ist, was er tut, sodass man schon nicht die besten Kategorien auswählen soll, sondern sich dann vielleicht schon so eher in der Mitte orientiert."

In der Fachliteratur der empirischen Sozialforschung wird diese Vorgehensweise mit *Priming* – die Verknüpfung eines kognitiven Reizes mit beabsichtigten Handlungsimperativen (vgl. Furnham/Boo 2011; Tversky/Kahneman 1974) respektive *Anchoring* – die Angabe eines Orientierungspunktes, der in einer systematischen Verzerrung hin zum Referenzpunkt mündet (vgl. Kahneman et al. 2006; Smith/Kendall 1963) – bezeichnet. Beide Formen der Beeinflussung bilden elementare Fehler in der Konstruktion quantifizierender Erhebungsinstrumente.

Auch die Rezensionen, die neben dem quantifizierenden Rating als zweites Erhebungsinstrument im Rahmen von RADAR eingesetzt werden, erzeugen Zerrbilder: Mithilfe von offenen Fragen wird gewöhnlich die Zustimmung oder Ablehnung eines Ratings gezielt hinterfragt, um konkret und konstruktiv auf Bewertungen eingehen zu können. Jedoch dienen bei RADAR auch die qualitativen Beurteilungen vor allem dazu, der Arbeitgeberseite Rechtfertigungen für mögliche Sanktionen zu liefern, die von Vorgesetzten mittel- oder langfristig gegen Arbeitnehmer*innen eingesetzt werden können. Alle Beurteilungen – auch die der Schwächen – müssen stichhaltig argumentativ begründet und stets mit Beispielen belegt werden. Nur die Hälfte der abgegebenen Beurteilungen dürfen positiv (Stärken) sein, die andere Hälfte der Bewertungen muss sich zwingend mit Kritik (Schwächen) auseinandersetzen, wodurch künstlich erzwungene Bekenntnisse entstehen, wie dieses Zitat beispielhaft verdeutlicht:

> „I mean for example, if I think that someone really didn't have three development areas, I might think they only had one, like I still have to pick three. […] I don't have any prescriptive to say what could be done differently, but I think because of some of the rigidity of the what you need to do, there might be some confusion with the feedback."

Die Struktur des Systems ist folglich so angelegt, dass sie das Management kontinuierlich mit Beispielen und Argumenten versorgt, die gegen Beschäftigte in Anschlag gebracht werden können.

Unter dem Strich lassen sich Ratingsysteme wie RADAR keineswegs als Instrumente zur Stärkung des Einflusses der Beschäftigten, zur Steigerung der Objektivität von Leistungsbeurteilungen und zur Erhöhung der Transparenz im Unternehmen beschreiben. Vielmehr handelt es sich zumindest bei dem von uns untersuchten Fall um ein hochgradig vermachtetes System, das einseitig den Interessen des Managements dient, indem es die Selbstkontrolle der Beschäftigten antizipiert und scheinobjektive Daten erzeugt, die zur Legitimierung von Managemententscheidungen eingesetzt werden können.

3.2 Arbeitsqualität und Coping

Die beschriebenen Zusammenhänge sind aus Sicht unserer Interviewees mit erheblichen Konsequenzen für das Betriebsklima und die Arbeitsqualität verbunden. Ein zentraler Kritikpunkt der Beschäftigten betrifft dabei die mit dem System implementierte Diffusion von Verantwortung im Unternehmen. Sie entsteht durch die unklare Zuordnung von Verantwortung im Rahmen von RADAR. Da die Kritik der Kolleg*innen anonym abgegeben wird und Vorgesetzte sich letztlich auf die Macht dieser nackten, nivellierenden Zahlen stützen, ist unklar, wem überhaupt eine Verantwortung für etwaige Sanktionen zugeschrieben werden kann. Dies wird von vielen unserer Interviewten beklagt:

> „Gefühlt ist es so eine dezentralisierte Entscheidung, es gibt keine Person mehr, die irgendwie was konkret entschieden hat. Also meine Mitarbeiter und Kolleginnen entscheiden praktisch mit über mich. Also, wenn irgendwas nicht stimmt und ich mich beschweren will, habe ich nicht meinen Chef, wo ich hingehe und sagen kann: ‚Nein, sehe ich anders und die Entscheidung finde ich nicht richtig' und kann mit einer Person diskutieren, sondern das ist komplett verschwommen."

Der bereits beschriebene Leistungsdruck und die Verunsicherung aufgrund der permanenten wechselseitigen Beobachtung führen letztlich auf der handlungspraktischen Ebene zu diversen Praktiken des passiven Widerstands gegen RADAR.

Solche Praktiken stehen aus unserer Sicht für nicht-intendierte Effekte des Systems, die deutlich von den Zielen abweichen, die das Management mit RADAR verfolgt. Wir deuten diese Praktiken als Versuche der Beschäftigten, mit den neuen Belastungen umzugehen (Coping), die im Rahmen des neuen Systems durch Veränderungen in der Arbeitssituation entstehen, und folglich als einen Effekt einer mit RADAR erfolgten Verschlechterung der Arbeitsqualität.

Zwei der am häufigsten genannte Praktiken, deren Einordnung zwischen Coping und Widerstand nicht ganz leicht fällt, bestehen erstens darin, positive Arbeitserfahrungen absichtlich zu überzeichnen, und zweitens darin, die eigenen sozialen Netzwerke im Unternehmen strate-

gisch auszuschöpfen, um den Bewertungsalgorithmus zu unterwandern. Beschäftigte geben beispielsweise an, „immer einen Ticken besser“ zu bewerten und Ratings von Kolleg*innen einzuwerben, von denen sie eine gute Bewertung erwarten:

> „Ich hab’ natürlich diejenigen genommen, bei denen ich wusste, okay, die haben mich eingearbeitet, dort vor allem diejenigen, bei denen mir klar war, dass sie von meiner Arbeit irgendwie überzeugt waren. Und ich hab’ bewusst Leute rausgelassen, von denen ich dachte, okay, möglicherweise hat jetzt derjenige mal irgendwie einen Fehler von mir gesehen.“

Um diesem Nutzerverhalten entgegenzutreten, ist es seit RADAR 2.0 verpflichtend, dass in den Evaluationen ebenso viele Schwächen wie Stärken bewertet werden; wenn beispielsweise sechs Themenbereiche zu einer Person beurteilt werden, ist die Angabe von wenigstens drei Schwächen bindend, bei acht evaluierten Themenbereichen sind es sogar mindestens vier Schwächen etc. Gleichzeitig müssen die *Werte* im Rating der Schwächen *geringer ausfallen* als bei den Stärken, sodass Schwächen nicht mit einem allzu guten Rating versehen werden können. Damit soll – nach unserer Einschätzung – verhindert werden, dass die Befragten die Bewertung von Schwächen durch ein gutes oder sehr gutes Rating (hohe Werte) relativieren. Um dies zu verhindern, greift der Algorithmus ein: Sobald beispielsweise bei wenigstens einem Rating der Stärken nur drei von sechs möglichen Punkten vergeben wurden, lässt die Software bei sämtlichen Ratings der Schwächen nur noch die Vergabe von maximal drei Punkten zu. Außerdem müssen Ratings mit zunehmender Wertigkeit mit entsprechend hochwertigen Argumenten und Beispielen belegt werden, die den von der Skalenbeschriftung suggerierten Anspruch erfüllen. Dabei nehmen einerseits die Abstände und wahrgenommenen Hürden von Skalenpunkt zu Skalenpunkt nicht gleichmäßig, sondern eher schneeballartig zu; andererseits werden Beurteilungen prinzipiell an einem geringen Anker- bzw. Ausgangswert der Ratingskala fixiert, dessen kontraintuitive Beschriftung zudem von vornherein eine Verzerrung *(Bias)* erzeugt: Beispielsweise ist bei der sechsstufigen Skala bereits die dritte Stufe mit „strong, consistently meets expectations“ und mit „company standard“ beschriftet. Mit den drei Stufen rechts von diesem Ankerpunkt sollen also weit über dem Durchschnitt stehende Leistungen assoziiert werden, wobei die als „Unicorn-level“ bezeichnete sechste Stufe ohnehin nur in absoluten Ausnahmefällen in Betracht kommen dürfte.

Sehr kurz vor der Periode der umfassenden Evaluationen können die Beschäftigten zunächst selbstständig mindestens zehn Evaluationsberechtigte vorschlagen, von denen die direkten Vorgesetzten letztlich *fünf* auswählen. Der späte Zeitpunkt dieser Nominierungsphase verhindert, dass nur eine geringe Anzahl an passenden personenspezifischen Beispielen für die Leistungsbeurteilung gesammelt wird. Stattdessen wird während der *gesamten* Zeit vor der Evaluationsperiode – eher zufällig und damit breit gestreut – beobachtet und gesammelt, womit ein permanentes und weitestgehend alle Personen umfassendes Beobachtungssystem realisiert wird.

3.3 Informationsasymmetrien und Lohnkontrolle

Ein großer Teil des Ärgers über RADAR geht, wie bereits erwähnt, auf die Intransparenzen des Systems zurück. Die Unwissenheit über die Funktionen der Technologie verstärkt dabei das Gefühl der Machtlosigkeit bei den Beschäftigten. Weder wurden die Funktionen des Systems in erklärenden Einweisungen kritisch erörtert, noch wurde die Logik offengelegt, nach der die erhobenen Daten aggregiert und verarbeitet werden. Wie der zum System gehörende Algorithmus die individuellen Scores bildet, ist den Beschäftigten vollkommen unbekannt. Auch der Verbleib der eigenen Daten und ihre Nutzung jenseits der – zumindest teilweise – bekannten Verwendungsweisen bleiben unklar. RADAR wird damit zu mehr als einer Software zur Erhebung von Daten. Es bildet, so fürchten jedenfalls einige Beschäftigte, den Grundstein einer digitalen Form der Personalakte, in der Informationen aus einer Vielzahl von Quellen gesammelt, dauerhaft gespeichert und jederzeit sanktionsrelevant werden können.

Diese strukturellen Informationsasymmetrien haben Folgen für die betriebliche Sozialstruktur. Kehren wir noch einmal zur Struktur der Rating-Technologie selbst zurück: Sie ist in spezifischer Weise vom Management gestaltet. Durch die in der Erhebungsmethode angelegte systematische Verzerrung wird das Ergebnis des Scorings, die berechneten Leistungsbewertungen der Angestellten, planmäßig niedrig gehalten. Dies geschieht in dem Bewusstsein, dass die Ratings der Klassifikation der Mitarbeiter*innen dienen und in der Folge sanktionsrelevant sind. Dies betrifft vor allem Entgeltfragen, die über die jeweiligen Scores vorstrukturiert werden. Letzten Endes dient die Einteilung der Belegschaft in drei Gruppen (Low-, Good- und Top-Performer) nicht nur der *Zuteilung von Aufstiegsoptionen* in der betrieblichen Hierarchie. Sie ist auch an die *Lohnstruktur* des Unternehmens gekoppelt. Während Top-Performer sich für Lohnsteigerungen qualifizieren, legt eine *exemplarische* Analyse der Gehaltssteigerung von Good-Performern nahe, dass sich deren Lohnzuwächse lediglich im Bereich des Inflationsausgleichs bewegen. Low-Performer, so zumindest die Beobachtung unserer Gesprächspartner, müssten nicht nur fürchten, dass bei noch nicht erfolgter Entfristung ihre Stelle nicht verlängert werde. Sie erhielten in der Regel auch keine Lohnsteigerungen, was in der Praxis sogar Reallohnverluste durch Inflation bedeuten würde.

Einzelne Beschäftigte erkennen durchaus die zu Grunde liegende Logik des Systems, das eben nicht nur auf Arbeitsdisziplin, sondern insbesondere auf Lohnrepression ausgerichtet ist und diese mittels „Legitimation durch Verfahren“ (Luhmann 2001 [1969]) objektiviert:

> „Ich bin der Meinung, dass es halt der Kontrolle dient, und irgendwie habe ich auch den Eindruck, dass es dann so eine Art Rechtfertigung im Endeffekt dafür ist, also für diejenigen, die Gehaltsentscheidungen dann treffen im Folgenden, dass man keine Gehaltserhöhung zahlen muss […] schon allein dadurch, dass man halt nie das Non-Plus-Ultra oder irgendwie die höchste Stufe auf dieser Bewertungsskala anwählen wird."

Besonders interessant ist nun, wie groß die Anteile der Belegschaft sind, die jeweils auf die drei Gruppen entfallen. Wir haben hierzu nur inkonsistente Indizien. Unseren Gesprächspartner*innen zufolge ist der Anteil an Top-Performern in den meisten Abteilungen auf 2 bis 3 % der Beschäftigten begrenzt. Wichtiger als die Frage, ob diese Zahl korrekt ist, scheint uns jedoch der Umstand, dass das Management die relativen Anteile der Leistungsgruppen im Grunde nach Belieben festlegen kann, um beispielweise die Lohnkosten zu deckeln, da es den Sortierungsalgorithmus kontrolliert. Die Basis hierfür sind die dem System eingeschriebenen Informationsasymmetrien. In tentativer Generalisierung lässt sich daraus schließen, dass digitale Ratingsysteme dem Management verlockende Optionen der Lohnkontrolle bieten, die mit der Macht *vermeintlich* objektiver Zahlen aus den Bewertungen der Beschäftigten selbst legitimiert werden können.

4 Fazit und Implikationen

Horizontale Digitalratings sind eine zentrale Komponente jüngerer Strategien des algorithmischen Managements. Von Unternehmensseite werden sie – nicht nur im Untersuchungsfall – als Instrumente zur Erhöhung der Transparenz von Personalentscheidungen, als Mittel zur Steigerung der Motivation von Beschäftigten und als datengetriebene und somit als Objektivität verheißende Vehikel zur Leistungsbeurteilung beworben. Der von uns beschriebene Fall eines besonders ambitionierten Ratingsystems zeigt, dass die verwendeten Technologien sehr einseitig im Unternehmensinteresse wirken und mit spezifischen Effekten für die Arbeitssituation verbunden sein können.

Systeme wie RADAR können *erstens* eine Intensivierung betrieblicher Kontrolle zur Folge haben. Da sich in Worker-Coworker-Ratings vor allem Beschäftigte der gleichen Hierarchiestufe gegenseitig bewerten, wird das Gefühl der Leistungskontrolle ubiquitär. Die Teilnehmer*innen der Ratings stehen zudem in einem Konkurrenzverhältnis um knappe betriebliche Aufstiegschancen, was nicht nur das Vertrauen in das System selbst untergräbt, sondern auch das Gefühl der Kontrolle verstärkt. Ferner sind Ratings als Kontrolltechnologien in die hochgradig vermachtete Beziehung zwischen Management und Beschäftigten eingelassen. Die Kontrolle des Managements über das Design und die Anwendung von Ratingtechnologien ermöglicht das systematische Ausnutzen von Informationsasymmetrien, wie es sich am Beispiel RADAR in Form diverser Intransparenzen sowie der im Design angelegten Präjustierung von Ergebnissen *(Priming, Anchoring, Kontrolle der Auswertung/Gewichtung)* zeigt.

Zweitens können mit derartigen Systemen nicht-intendierte Effekte verbunden sein, die womöglich auch den Nutzen horizontaler Ratings für die Unternehmensseite konterkarieren. Im Untersuchungsfall wenden die Beschäftigten ein beachtliches Maß an Energie dafür auf, sich gegen schlechte Bewertungen abzusichern, und müssen relevante Teile ihrer Arbeitszeit für das Einpflegen von Daten verwenden. Ob die durch das System erzeugte Selbstdisziplinierung diese nicht-intendierten Effekte wettmacht, kann niemand wissen. Die Arbeitsqualität leidet im Untersuchungsfall jedenfalls deutlich, was mit negativen Implikationen für die Motivation der Beschäftigten verbunden ist – wo laut Unternehmensdarstellung doch gerade der gegenteilige Effekt intendiert wird.

Drittens scheint uns der Anspruch, Karriereoptionen durch die dem System entspringenden Scores berechenbar zu machen, durchaus einlösbar – allerdings nur für das Management. Während die Beschäftigten – jedenfalls im Untersuchungsfall – nur über wenig Wissen verfügen, wie sie ihre eigenen Score verbessern können, kontrolliert das Management neben dem Design des Instruments auch die Datenauswertung. Es kann daher im Prinzip festlegen, wie groß die jeweiligen Gruppen (Low-, Good- und Top-Performer) bei RADAR sein sollen. Wird nun die Lohnentwicklung an die Zugehörigkeit zu einer dieser Gruppen gekoppelt, eröffnet dies dem Management die Möglichkeit, neben unternehmensinternen Aufstiegen auch die Löhne zu kontrollieren. Damit werden moderne Formen von Digitalratings potenziell zu Instrumenten der Herstellung neuer Wertigkeiten und erzeugen ebenfalls neue betriebliche Ungleichheiten.

Beim Vormarsch neuerer Digitalratings in der tertiären Arbeitswelt sind daher die Akteure der Mitbestimmung[5] gefragt. Das Entgelt betreffende Entwicklungen sind nach § 87 Abs. 1 Nr. 10 des Betriebsverfassungsgesetzes mitbestimmungspflichtig. Dies gilt auch für „technische Einrichtungen", die zur Kontrolle von Beschäftigten „bestimmt sind" (§ 87 Abs. 1 Nr. 6 BetrVG). Ein großes Problem in diesem Zusammenhang stellt die Komplexität der Systeme und gegebenenfalls ihr proprietärer Charakter dar.[6] Ratings stehen daher für einen im Kontext des

5 Betriebliche Mitbestimmung nach dem BetrVG kann freilich nur dort wirksam ausgeübt werden, wo Betriebsräte tatsächlich existieren. Gerade in großen Teilen des Dienstleistungssektors stellt dies eine hohe Hürde dar.

6 Kommen etwa Instrumente von Drittanbietern zum Einsatz, muss geprüft werden, inwiefern eine größere Transparenz der jeweiligen Anwendungen durchsetzbar ist.

algorithmischen Managements steigenden Bedarf an Schulungen für Betriebsrät*innen und Gewerkschafter*innen zu den Funktionslogiken etwaiger Systeme.[7] Um Transparenzforderungen durchsetzbar zu machen, mag ebenso das neue Datenschutzrecht ein aussichtsreicher Ansatz sein, vor allem der vieldiskutierte Artikel 12 der Datenschutzgrundverordnung (DSGVO), in dem die Transparenz von Information, Kommunikation und Modalitäten geregelt wird. Dort heißt es im Erwägungsgrund 58: „Der Grundsatz der Transparenz setzt voraus, dass eine für die Öffentlichkeit oder die betroffene Person bestimmte Information präzise, leicht zugänglich und verständlich sowie in klarer und einfacher Sprache abgefasst ist und gegebenenfalls zusätzlich visuelle Elemente verwendet werden." Diese allgemeine juristische Definition der Transparenz technischer Systeme wird überdies durch den Artikel 22 der DSGVO, welcher sich mit automatisierten Entscheidungen im Einzelfall einschließlich Profiling auseinandersetzt, weiter ausgeführt und direkt auf die algorithmische Kontrolle von Personen bezogen. Das Aufbrechen der Blackbox-Logik (vgl. Pasquale 2015) algorithmischer Arbeitskontrollanwendungen bildet jedenfalls die Voraussetzung für deren Gestaltung im Sinne guter Arbeit im Kontext der Mitbestimmung. ■

LITERATUR

Attwood-Charles, W. (2019): Technology and control. Institutional work and digital platforms, unveröffentlichtes Manuskript

Beverungen, A. (2017): Algorithmisches Management, in: Beyes, T. / Metelmann, J. / Pias, C. (Hrsg.): Nach der Revolution. Ein Brevier digitaler Kulturen, Berlin, S. 52–63

Bröckling, U. (2003): Das demokratisierte Panopticon. Subjektivierung und Kontrolle im 360-Feedback, in: Honneth, A. / Saar, M. (Hrsg.): Michel Foucault. Zwischenbilanz einer Rezeption. Frankfurter Foucault-Konferenz 2001, Frankfurt a. M., S. 77–93

Evans, L. / Kitchin, R. (2018): A smart place to work? Big data systems, labour, control and modern retail stores, in: New Technology, Work and Employment 33 (1), S. 44–57

Fleenor, J. W. / Prince, J. M. (1997): Using 360-degree feedback in organizations, Center for Creative Leadership, Greensboro NC

Friedman, A. (1987): Managementstrategien und Technologie: Auf dem Weg zu einer komplexen Theorie des Arbeitsprozesses, in: Hildebrand, E. / Seltz, R. (Hrsg.): Managementstrategien und Kontrolle. Eine Einführung in die Labour Process Debate, Berlin, S. 99–113

Furnham, A. / Boo, H. C. (2011): A literature review of the anchoring effect, in: The Journal of Socio-economics 40 (1), S. 35–42

Glaser, B. G. / Strauss, A. L. / Strutzel, E. (1968): The discovery of grounded theory; strategies for qualitative research, in: Nursing Research, 17 (4), S. 364

Hirsch-Kreinsen, H. / Ittermann, P. / Niehaus, J. (Hrsg.) (2015): Digitalisierung industrieller Arbeit. Die Vision Industrie 4.0 und ihre sozialen Herausforderungen, Baden-Baden

Kahneman, D. / Krueger, A. B. / Schkade, D. / Schwarz, N. / Stone, A. A. (2006): Would you be happier if you were richer? A focusing illusion, in: Science 312 (5782), S. 1908–1910

Kantor, J. / Streitfeld, D. (2015): Inside Amazon. Wrestling big ideas in a bruising workplace, http://www.nytimes.com/2015/08/16/technology/inside-amazon-wrestling-big-ideas-in-a-bruising-workplace.html (letzter Zugriff: 27. 04. 2016)

Kirchner, S. / Beyer, J. (2016): Die Plattformlogik als digitale Marktordnung. Wie die Digitalisierung Kopplungen von Unternehmen löst und Märkte transformiert, in: Zeitschrift für Soziologie 45 (5), S. 324–339

Lee, M. K. (2015): Working with machines. The impact of algorithmic and data-driven management on human workers, https://www.researchgate.net/publication/277875720_Working_with_Machines_The_Impact_of_Algorithmic_and_Data-Driven_Management_on_Human_Workers (letzter Zugriff: 13. 8. 2019)

Lee, M. K. / Baykal, S. (2017): Algorithmic mediation in group decisions. Fairness perceptions of algorithmically mediated vs. discussion-based social division, https://www.researchgate.net/publication/313738865_Algorithmic_Mediation_in_Group_Decisions_Fairness_Perceptions_of_Algorithmically_Mediated_vs_Discussion-Based_Social_Division (letzter Zugriff: 13. 08. 2019)

Luhmann, N. (2001 [1969]): Legitimation durch Verfahren, 6. Aufl., Frankfurt a. M.

Marrs, K. (2018): Herrschaft und Kontrolle in der Arbeit, in: Böhle, F. / Voß, G. G. / Wachtler, G. (Hrsg.): Handbuch Arbeitssoziologie, 2. Aufl., Wiesbaden, S. 473–502

Moore, P. V. (2018a): The quantified self in precarity. Work, technology and what counts, Abingdon

Moore, P. V. (2018b): Tracking affective labour for agility in the quantified workplace, in: Body & Society 24 (3), S. 39–67

Moore, P. V. / Piwek, L. (2017): Regulating wellbeing in the brave new quantified workplace, in: Employee Relations 39 (3), S. 308–316

Pasquale, F. (2015): The black box society. The secret algorithms that control money and information, Cambridge MA

Schaupp, S. / Staab, P. (2018): Rekursivität und Horizontalisierung. Das kommerzielle Internet als Vorbild digitalisierter Arbeit, in: Arbeits- und Industriesoziologische Studien 11 (2), S. 294–307

Schor, J. B. / Attwood-Charles, W. (2017): The 'sharing' economy: Labor, inequality, and social connection on for-profit platforms, in: Sociology Compass 11 (8), https://doi.org/10.1111/soc4.12493

Schor, J. B. / Attwood-Charles, W. / Cansoy, M. / Ladegaard, I. / Wengronowitz, R. (2017): Dependence and precarity in the platform economy, https://www.bc.edu/content/dam/files/schools/cas_sites/sociology/pdf/Dependence%20and%20Precarity%20Feb%202017.pdf (letzter Zugriff: 15. 06. 2019)

Smith, P. C. / Kendall, L. M. (1963): Retranslation of expectations. An approach to the construction of unambiguous anchors for rating scales, in: Journal of Applied Psychology 47 (2), S. 149–155

Staab, P. (2019): Digitaler Kapitalismus. Markt und Herrschaft in der Ökonomie der Unknappheit, Berlin

Tversky, A. / Kahneman, D. (1974): Judgment under uncertainty: Heuristics and biases, in: Science, 185 (4157), S. 1124–1131

AUTOREN

PHILIPP STAAB, Dr., Professor für die Soziologie der Zukunft der Arbeit an der Humboldt Universität zu Berlin und am Einstein Center Digital Future (ECDF). Forschungsschwerpunkte: Arbeitssoziologie, politische Ökonomie und Technik-soziologie.

@ philipp.s.staab@hu-berlin.de

SASCHA-CHRISTOPHER GESCHKE studiert Sozialwissenschaften an der Humboldt Universität zu Berlin und ist Mitarbeiter am Institut für Sozialwissenschaften (ISW) und am Deutschen Institut für Wirtschaftsforschung (DIW). Arbeitsschwerpunkte: Digitalisierung, Arbeitsmarkt-, Migrations- und Lebenslaufforschung, Methoden der empirischen Sozialforschung.

@ s.geschke@hu-berlin.de

7 Wie wird beispielsweise die Datenspeicherung und -verarbeitung geregelt? Was ist hier zulässig? Welche Löschfristen werden gesetzt? Welche Zweckbindung der erhobenen Daten wird definiert (eine Information hierüber bildet die notwendige Voraussetzung, um deren Einhaltung überhaupt prüfen zu können)?