# GMD Report 60

GMD –
Forschungszentrum
Informationstechnik
GmbH

Michael May,
Uwe Oestermeier (Hrsg.)

# KogWis99
# Workshop „Kausalität“

4. Fachtagung der Gesellschaft

für Kognitionswissenschaft

an der Universität Bielefeld

28. September - 1. Oktober 1999

September 1999

© GMD 1999

GMD –
Forschungszentrum Informationstechnik GmbH
Schloß Birlinghoven
D-53754 Sankt Augustin
Germany
Telefon +49 -2241 -14 -0
Telefax +49 -2241 -14 -2618
http://www.gmd.de

In der Reihe GMD Report werden Forschungs- und Entwicklungsergebnisse aus der GMD zum wissenschaftlichen, nichtkommerziellen Gebrauch veröffentlicht. Jegliche Inhaltsänderung des Dokuments sowie die entgeltliche Weitergabe sind verboten.

The purpose of the GMD Report is the dissemination of research work for scientific non-commercial use. The commercial distribution of this document is prohibited, as is any modification of its content.

**Anschriften der Herausgeber/Addresses of the editors:**

Dr. Michael May
Institut für Autonome intelligente Systeme
GMD – Forschungszentrum Informationstechnik GmbH
D-53754 Sankt Augustin
E-mail: Michael.May@gmd.de

Dr. Uwe Oestermeier
Abteilung für Angewandte Kognitionswissenschaft
Deutsches Institut für Fernstudienforschung
an der Universität Tübingen
Konrad-Adenauer-Straße 40
D-72072 Tübingen
E-mail: uwe_oestermeier@diff.uni-tuebingen.de

ISSN 1435-2702

## Abstract

The category of causation is fundamental for human cognition. Its investigation has a long tradition in many disciplines related to cognitive science: cognitive psychology investigates the mental processes of causal reasoning and judgment. Epistemology and philosophy of science ask for a justification for taking causal hypothesis as guide for future action and they try to explicate the notion of causation; artificial intelligence addresses causal models as a foundation for robust inference systems. The workshop presented and discussed current research in these disciplines. It took place at 30.9.1999 in Bielefeld. This book contains the contributions presented at the workshop.

**Keywords:** Causation, Causal Reasoning, Cognitive Science

## Zusammenfassung

Die Kategorie der Kausalität ist grundlegend für das menschliche Denken. Entsprechend hat ihre Untersuchungen in vielen kognitionswissenschaftlich relevanten Disziplinen eine lange Tradition: die kognitive Psychologie untersucht die mentalen Prozesse des kausalen Schließens und Urteilens; die Philosophie und Wissenschaftstheorie fragt nach der Berechtigung, mit der wir kausale Hypothesen als Grundlage für zukünftiges Handeln nehmen und versucht den Begriff der Kausalität zu explizieren; die Künstliche Intelligenz nimmt Kausalmodelle als Grundlage für robuste Inferenzsysteme. In dem Workshop wurden aktuelle Forschungsergebnisse aus diesen verschiedenen Disziplinen vorgestellt und diskutiert. Der Workshop fand am 30.9.1999 in Bielefeld statt. Der vorliegende Band enthält die eingereichten Workshopbeiträge.

**Schlagworte:** Kausalität, Kausales Schließen, Kognitionswissenschaft

# Inhalt

Michael May & Uwe Oestermeier: Einleitung ..... 7

Heiko Hecht & Dirk Kerzel, Universität Bielefeld und Max-PIanck-Institut für Psychologische Forschung: Die Wahrnehmung von Kausalität ..... 12

Marc Buehner, Universität Sheffield: How do humans induce causation from covariation? ..... 19

Marcus Plach, Universität des Saarlandes: Bayessche Netze als Modelle der Repräsentation von Kausalwissen ..... 31

York Hagmayer & Michael R. Waldmann, Universität Göttingen: Sind statistische Verfahren zur Kausalanalyse adäquate Modelle kausalen Denkens? ..... 39

Christof Nachtigall, Rolf Steyer & Olivia Wüthrich-Martone, Universität Jena: Kausale Effekte in der empirischen Sozialforschung ..... 44

Uwe Oestermeier, DIFF Tübingen: Die Mathematisierung der Wissenschaften und die Beobachtbarkeit von Kausalrelationen ..... 50

Michael May, GMD Sankt Augustin: Die Kontextabhängigkeit kausaler Erklärungen ..... 58

# Workshop Kausalität auf der KogWis99

Michael May

GMD-Forschungszentrum Informationstechnik,
Institut für Autonome intelligente Systeme, Knowledge Discovery Team
Schloß Birlinghoven, D-53754 Sankt Augustin
Tel. +49-2241-14-2039, email: michael.may@gmd.de

Uwe Oestermeier

Abteilung für Angewandte Kognitionswissenschaft
Deutsches Institut für Fernstudienforschung an der Universität Tübingen
Konrad-Adenauer-Str. 40, D-72072 Tübingen
Tel. +49-7071-979-208, email: uwe_oestermeier@diff.uni-tuebingen.de

Die Kategorie der Kausalität ist grundlegend für das menschliche Denken. Entsprechend hat ihre Untersuchungen in vielen kognitionswissenschaftlich relevanten Disziplinen eine lange Tradition: die kognitive Psychologie untersucht die mentalen Prozesse des kausalen Schließens und Urteilens; die Philosophie und Wissenschaftstheorie fragt nach der Berechtigung, mit der wir kausale Hypothesen als Grundlage für zukünftiges Handeln nehmen und versucht den Begriff der Kausalität zu explizieren; die Künstliche Intelligenz nimmt Kausalmodelle als Grundlage für robuste Inferenzsysteme. In dem Workshop wurden aktuelle Forschungsergebnisse aus diesen verschiedenen Disziplinen vorgestellt und diskutiert. Der Workshop fand am 30.9.1999 in Bielefeld statt. Der vorliegende Band enthält die eingereichten Beiträge.

## Zusammenfassung der Beiträge

**Heiko Hecht & Dirk Kerzel, ZIF Bielefeld & Max-Planck-Institut für psychologische Forschung, München: Die Wahrnehmung von Kausalität**
Je komplexer vom Menschen zu beherrschende künstliche Systeme werden, desto mehr Aufmerksamkeit muß der Lücke zwischen wahrgenommener und tatsächlicher Kausalität in solchen Systemen geschenkt werden. Dieser Beitrag versucht diese Lücke systematisch offenzulegen und Lösungsvorschläge für ihre Schließung vorzustellen, und zwar aus wahrnehmungspsychologischer Sicht. Eine phänomenologische Analyse der Wahrnehmung von Kausalität ist erstmals von Heider und Simmel (1944) vorgenommen und dann von Michotte (1954) und Houssiadas (1964) ergänzt worden. Sie ist bis heute populär und geht davon aus, daß wir einen direkten perzeptuellen Zugang zu Wirkzusammenhängen haben. So sehen wir beispielsweise an einer Bewegung, ob diese ursprünglich selbstverursacht oder von außen angestoßen wurde. Eigene Untersuchungen haben allerdings gezeigt, daß unsere Wahrnehmungsfähigkeit kausaler Strukturen an ganz enge Grenzen gebunden ist. So nehmen wir etwa ballistische Bewegungen bis zur Umkehrung der Kausalverhältnisse verzerrt war. Wir stellen einen Erklärungsansatz vor, der unsere Wahrnehmungsgrenzen aus der eigenen Körpermotorik heraus begründet. Die Biodynamik unserer Effektoren gibt den Rahmen dessen vor, welche Kausalbeziehungen perzeptuell lernbar sind.

**Marc Buehner, University of Sheffield: How do humans induce causation from covariation?**
Three approaches are experimentally contrasted: the classic contingency model, the Rescorla-Wagner (1972) model, and the power PC theory (Cheng, 1997). The approaches differ in their predictions for situations in which $\Delta P$ (the difference between the probability of an effect $e$ in the presence of a candidate cause $c$ and that in its absence) is constant, but the base rate of $e$ (the probability of $e$ given the absence of $c$) varies. The experiments reported here demonstrate a substantial influence of this

base-rate across conditions in which $\Delta P$ remained constant. This influence was diametrically opposite for generative candidate causes as for preventive candidates, as predicted uniquely by the power PC theory. Purely covariation-based approaches, including a benchmark associative learning theory, do not explain these opposite influences without additional empirical assumptions.

**Marcus Plach, Universität des Saarlandes: Bayessche Netze als Modelle der Repräsentation von Kausalwissen**

Bei der kognitionspsychologischen Untersuchung von Prozessen der Wissensaktualisierung (*belief-updating*) wurde in der Vergangenheit die Bedeutung von Kausalannahmen für die Evaluation neuer Information vernachlässigt. In der vorliegenden Arbeit wird untersucht, inwieweit sogenannte ‚Bayessche Netze' ein plausibles Modell der kognitiven Repräsentation von Annahmen über probabilistische Kausalzusammenhänge und der darauf operierenden Inferenzprozesse darstellen. Es wird zunächst die Beziehung zwischen verschiedenen kausalen Grundmodellen und Abhängigkeitsstrukturen von Ereignissen, wie sie in Bayesschen Netze repräsentiert werden können, analysiert. Im Anschluß werden Daten aus zwei Experimenten zur Revision von Überzeugungsstärken und Vergleiche mit den Vorhersagen eines Bayesschen Netzes vorgestellt. Implikationen für die Modellierung der Repräsentation und Verarbeitung von Kausalwissen werden diskutiert.

**York Hagmayer & Michael R. Waldmann, Universität Göttingen: Sind statistische Verfahren zur Kausalanalyse adäquate Modelle kausalen Denkens?**

In den letzten Jahrzehnten sind eine Reihe von Verfahren entwickelt worden, mit denen Kausalanalysen durchgeführt werden können. Am weitesten verbreitet sind lineare Strukturgleichungsmodelle wie LISREL. Diese Verfahren testen angenommene Kausalmodelle anhand der von ihnen implizierten Korrelationen zwischen den Variablen. Psychologische Forschung zum kausalen Denken wiederum hat gezeigt, daß Personen bei der Bestimmung kausaler Zusammenhänge auf statistische Zusammenhänge zurückgreifen und zwischen verschiedenen Kausalstrukturen unterscheiden. Wie Personen verschiedene Modelle gegeneinander testen, ist allerdings noch weitestgehend unklar. Strukturgleichungsmodelle machen genaue Angaben, welche Daten zu ermitteln und welche Zusammenhänge für eine Überprüfung eines Kausalmodells zu untersuchen sind. Sie bieten daher zumindest einen normativen Vergleichsmaßstab, an dem die Einschätzungen von Personen gemessen werden können. In einer Reihe von Experimenten sind wir nun der Frage nachgegangen, ob sie selbst Modell für das Denken sind.

**Christof Nachtigall, Rolf Steyer & Olivia Wüthrich-Martone, Universität Jena: Kausale Effekte in der empirischen Sozialforschung**

Ein wichtiger Schritt zur Präzisierung des Kausalitätsbegriffs ist das Konzept des individuellen und durchschnittlichen kausalen Effekts. Im Kontext der Psychologie spricht man von einem individuellen kausalen Effekt (ICE) als dem Effekt, den ein Treatment auf eine Person hätte im Vergleich zur Nichtbehandlung. Allerdings besteht in der Regel nicht die Möglichkeit, individuelle kausale Effekte zu bestimmen, denn man kann eine Person nicht gleichzeitig behandeln und nicht behandeln. Unter Umständen ist es jedoch möglich, einen durchschnittlichen kausalen Effekt (ACE) zu ermitteln. Dies ist der Durchschnitt aus den individuellen kausalen Effekten über alle Personen der betrachteten Population. Er beschreibt die mittlere Veränderung aufgrund des Treatments im Vergleich zur Kontrollgruppe. In dem Vortrag wird es darum gehen, wann und auf welche Weise Aussagen über durchschnittliche kausale Effekte gemacht werden können.

**Uwe Oestermeier, DIFF Tübingen: Die Mathematisierung der Wissenschaften und die Beobachtbarkeit von Kausalrelationen**

Seit langem bestehen in Philosophie, Wissenschaftstheorie und Psychologie unterschiedliche Auffassungen darüber, ob Kausalrelationen im Einzelfall beobachtbar sind oder nicht. Dieser Dissens läßt sich auflösen, wenn man davon ausgeht, dass in der Debatte unterschiedliche alltagssprachliche und normative Beobachtungsbegriffe Verwendung finden. Es wird gezeigt, dass ein normativer Beobachtungsbegriff, der strikt zwischen Beobachtung und kausaler Interpretation trennt sowie die

Intersubjektivität, Objektivität und Präzision von Beobachtungen fordert, kognitionspsychologisch gesehen Anforderungen an das Erkenntnissubjekt stellt, die in der Praxis erst mit der Mathematisierung der Wissenschaften erfüllt werden konnten. Diese Mathematisierung wird als eine Erweiterung kognitiver Fähigkeiten durch externe Repräsentationen und kognitive Artefakte beschrieben.

**Michael May, GMD, Sankt Augustin: Die Kontextabhängigkeit kausaler Erklärungen**
Folgendes Beispiel wird häufig benutzt, um die Kontextabhängigkeit kausaler Erklärungen herauszustellen: Der Arzt führt als Ursache für den Tod des Fahrers starke innere Blutungen an, der Anwalt unachtsames Fahren, der Autokonstrukteur einen Fehler in der Konstruktion der Bremsen. Was der jeweilige Sachverständige als *die Ursache* herausgreift, ist abhängig von seiner Fragestellung und einem pragmatisch gegebenen Kontext. Hieraus wird manchmal der radikale Schluß gezogen, daß es sinnlos ist, von Kausalzusammenhängen zu reden, welche unabhängig von unseren Interessen und unserem Vorwissen bestehen. Ich werde ein Modell kausaler Erklärung vorstellen, welches der Relativität und Kontextabhängigkeit kausaler Erklärungen Rechnung trägt und dennoch eine formale Behandlung des Erklärungsbegriffs erlaubt. Es weicht in einem zentralen Punkt von den gängigen Erklärungsmodellen, wie sie in der Wissenschaftstheorie entwickelt wurden (Hempel 1965 etc., Salmon 1984, van Fraassen 1980 oder Gärdenfors 1988), ab: Eine kausale Erklärung wird nicht als potentielle deterministische oder statistische Prognose aufgefaßt, sondern als *Diagnose*, bei der aus dem Bestehen der Wirkung unter Ausschluß von Alternativursachen auf das Bestehen einer Ursache geschlossen wird.

**Gerd Graßhoff, Universität Bern: Instrumente, Forschungspraxis und Kausales Schließen**
Untersuchungen von kognitiven Prozessen bei Forschungsaktivitäten sind bislang größtenteils an historischen Fallstudien durchgeführt worden. Wiederholt wurde an solchen Studien kritisiert, daß die retrospektive Rekonstruktion der wissenschaftlichen Arbeitsweise zu einer Rationalisierung der Forschungshandlungen führt, bei der die tatsächlich befolgten kognitiven Prozesse nur unzureichend erfaßt werden. Um die Stichhaltigkeit solcher Einwände zu prüfen, wurden an einem synchron zur wissenschaftstheoretischen Untersuchung durchgeführten Forschungsprojekt aus der Astronomie Teile dieses Forschungsvorhabens von Wissenschaftstheoretikern selbst durchgeführt, protokolliert und methodologisch ausgewertet.
In Abwandlung des Begriffs vom teilnehmenden Beobachter war es das Ziel dieser Untersuchungen, durch beobachtende Teilnehmer an wissenschaftlichen Forschungsaktivitäten neue Evidenzen über die befolgten Forschungsstrategien zu gewinnen. Gegenstand des astronomischen Forschungsprojektes war die Entwicklung eines komplexen automatischen Klassifikationssystems zur Suche von bestimmten sternähnlichen Objekten auf der Basis digitalisierter Spektralaufnahmen. Vorgestellt wird, mit welchen Methoden die Untersuchung durchgeführt wurde, auf welche Weise komplexe Forschungsinstrumente entwickelt werden und wie dabei Regeln des kausalen Schließens angewendet werden.

# Ablauf des Workshops

## Donnerstag, 30. September 1999, 14:00 - 17:30 Uhr

| Zeit | Name(n) | Thema |
|---|---|---|
| 14:00 - 14:10 | Michael May & Uwe Oestermeier | Überblick und Einführung |
| 14:10 - 14:30 | Heiko Hecht & Dirk Kerzel | Die Wahrnehmung von Kausalität |
| 14:30 - 14:50 | Marc Buehner | How do humans induce causation from covariation? |
| 14:50 - 15:10 | Marcus Plach | Bayessche Netze als Modelle der Repräsentation von Kausalwissen |
| 15:10 - 15:30 | York Hagmayer & Michael R. Waldmann | Sind statistische Verfahren zur Kausalanalyse adäquate Modelle kausalen Denkens? |
| 15:30 - 15:45 | Pause | |
| 15:45 - 16:05 | Christof Nachtigall, Rolf Steyer & Olivia Wüthrich-Martone | Kausale Effekte in der empirischen Sozialforschung |
| 16:05 - 16:25 | Gerd Graßhoff | Instrumente, Forschungspraxis und Kausales Schließen |
| 16:25 - 16:45 | Michael May | Die Kontextabhängigkeit kausaler Erklärungen |
| 16:45 - 17:05 | Uwe Oestermeier | Die Mathematisierung der Wissenschaften und die Beobachtbarkeit von Kausalrelationen |
| 17:05 - 17:30 | Abschlußdiskussion | |

# Beiträge

# Die Wahrnehmung von Kausalität

**Heiko Hecht**
Zentrum für interdisziplinäre Forschung
Universität Bielefeld
E-mail: hecht@hrz.uni-bielefeld.de

**Dirk Kerzel**
Max-Planck-Institut für Psychologische Forschung,
München

## Abstract

Je komplexer vom Menschen zu beherrschende künstliche Systeme werden, desto mehr Aufmerksamkeit muß der Lücke zwischen wahrgenommener und tatsächlicher Kausalität in solchen Systemen geschenkt werden. Dieser Beitrag versucht diese Lücke systematisch offenzulegen und Lösungsvorschläge für ihre Schließung vorzustellen, und zwar aus wahrnehmungspsychologischer Sicht. Eine phänomenologische Analyse der Wahrnehmung von Kausalität ist erstmals von Heider und Simmel (1944) vorgenommen und dann von Michotte (1954) und Houssiadas (1964) ergänzt worden. Sie ist bis heute populär und geht davon aus, daß wir einen direkten perzeptuellen Zugang zu Wirkzusammenhängen haben. So sehen wir beispielsweise an einer Bewegung, ob diese ursprünglich selbstverursacht oder von außen angestoßen wurde. Eigene Untersuchungen haben allerdings gezeigt, daß unsere Wahrnehmungsfähigkeit kausaler Strukturen an ganz enge Grenzen gebunden ist. So nehmen wir etwa ballistische Bewegungen bis zur Umkehrung der Kausalverhältnisse verzerrt war. Ich stelle einen Erklärungsansatz vor, der unsere Wahrnehmungsgrenzen aus der eigenen Körpermotorik heraus begründet. Die Biodynamik unserer Effektoren gibt den Rahmen dessen vor, welche Kausalbeziehungen perzeptuell lernbar sind.

## Einleitung

Die Verbreitung von virtuellen Werkzeugen, von der Fernbedienung bis zur Telechirurgie, zeichnet sich durch eine immer größer werdende Lücke von erlebbaren Ursachen- und Wirkungsverhältnissen aus. Für die menschliche Wahrnehmung heißt dies, daß äußerlich sehr unähnliche Dinge in einen kausalen Zusammenhang gebracht werden müssen. Während für unsere Vorväter etwa ein mechanischer Hebel eine Weiche stellte, die Wirkursache über einen Dreh- und Hebelpunkt also unmittelbar wahrnehmbar war, ist heute nur noch per Mausklick ein farbiges Symbol auf dem Monitor zu manipulieren. Eine Fülle von weiteren Beispielen könnte angeführt werden, um zu untermauern, daß an unsere Kausalitätswahrnehmung zunehmend höhere Anforderungen gestellt werden, damit wir beim immer größer werdenden Abstand zwischen sichtbarer Ursache und sichtbarer Wirkung den Nexus weiterhin erfassen. Ich möchte hier die fest etablierte Meinung in Frage stellen, daß die Wahrnehmung von Kausalität unmittelbar und direkt vonstatten geht, ähnlich der Form- oder Farbwahrnehmung. Ich werde darlegen, daß Kausalitätswahrnehmung am ehesten mit der Aristotelischen Auffassung der Zweckursache verstanden werden kann. Daraus ergeben sich

weitreichende Konsequenzen für das Erlernen von Wirkzusammenhängen in künstlichen Umgebungen.

**Die Phänomenologie des Kausalen**

Seit David Hume konzentrierten sich Philosophen darauf, den Ursachenbegriff als Wirkursache zu verstehen. In diesem Sinne einer wissenschaftlichen Erklärung ist eine Ursache ein Ereignis, das hinreichend ist, um ein bestimmtes zweites Ereignis folgen zu lassen. So ist etwa der Aufprall einer bewegten Kugel auf eine ruhende Ursache dafür, daß letztere in Bewegung versetzt wird. Man kann dies ganz occasionalistisch sehen und bloße Koinzidenz bereits als Wirkursache verstehen, oder man kann zwischen Aufprall und angestoßener Bewegung noch weitere Mechanismen annehmen (Reibung, Energieübertragung, etc.), die das Kausalverhältnis weiter aufschlüsseln.

Für diese Form des Ursächlichen haben wir effiziente Wahrnehmungsmechanismen erworben. So hat etwa Albert Michotte (1954) nachgewiesen, daß Beobachter die Bewegungen zweier noch so simpler Objekte stets als kausal verursacht und aufeinander bezogen wahrnehmen. Setzt sich beispielsweise die "angestoßene" Kugel in Bewegung bevor sie überhaupt angestoßen wird, so ist dennoch der Eindruck unausweichlich, das die erste Kugel für die Bewegung der zweiten ursächlich verantwortlich ist. Michotte hat mit einer Drehscheibe experimentiert, die bis auf einen kleinen Schlitz von einer Abdeckung verdeckt war (siehe Abbildung Fig. 1). Die beiden kreisförmigen Linien werden dann durch den Schlitz betrachtet als sich annähernde und abstoßende Objekte erlebt. Der Eindruck der Kausalität ist sehr prägnant und erweist sich als äußerst stabil, d. h. er besteht auch wenn zeitliche und räumliche Abweichungen eingeführt werden. Z. B. wenn von zwei ruhenden Kugeln sich erst die eine und dann die andere in Bewegung setzt (in die gleiche Richtung ohne sie zu berühren), so entsteht der Eindruck als würde erstere die zweite hinter sich herziehen. Neuere komputerbasierte Methoden kommen, auch bei Kleinkindern, zum gleichen Ergebnis.

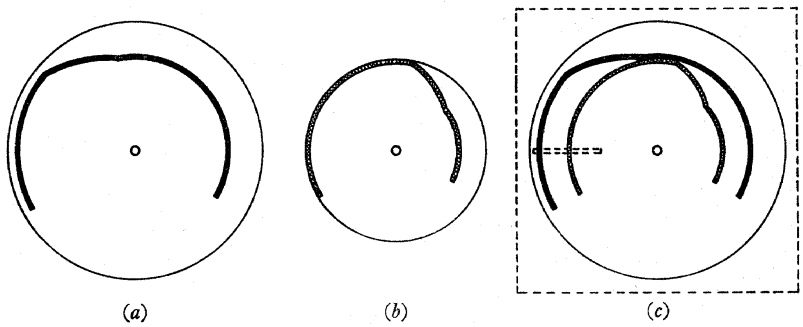


Fig. 1. Disks used by Michotte.

Dieser phänomenologische Tatbestand ist hinlänglich experimentell untersucht und variiert worden. Besonders bemerkenswert ist, daß erlebte Ursache und Wirkung in keinem direkten

räumlichen Kontakt stehen müssen. Houssiadas (1964) hat beispielsweise eine Vielzahl von Versuchen durchgeführt, in denen er räumliche und zeitliche Abstände einführte. In Abbildung 2 ist einer seiner Versuche schematisch dargestellt. Die Deformation des Rechtecks wird als vom Kreis verursacht erlebt, obwohl der Kreis nie das Rechteck berührt oder auch nur den Ort errreicht, den Teile des Rechtecks vorher eingenommen haben. Die Phasen 1 - 3 sind als zeitlich aufeinanderfolgend zu interpretieren.

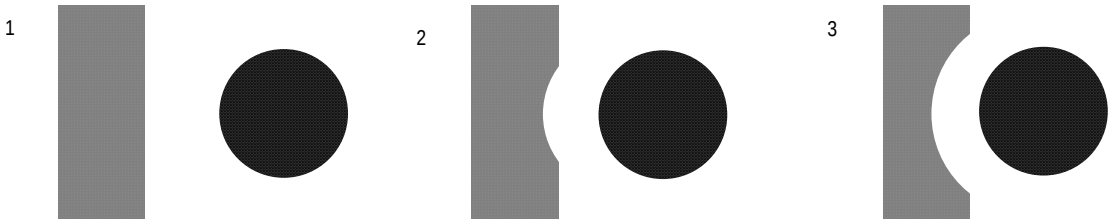



Abbildung 2

Neben diesem direkten, unmittelbaren Erlebnis von Kausalität steht die große Bereitschaft des visuellen Systems Wirkungen zu antizipieren. So haben etwa Hubbard (1995) und Bertamini (1993) gezeigt, daß die erlebte Wirkung eines sich bewegenden Objektes nicht nur von seiner momentanen Bewegungsrichtung abhängig ist, sondern Effekte eines Hindernisses (antizipierter Abprall) oder der Erdanziehungskraft automatisch in unsere Wahrnehmung einbezogen werden. Wir täuschen uns in der Position von Objekten immer in Richtung auf die antizipierte Bewegung, nie in die entgegengesetzte; und unsere Erinnerungsfehler sind im Einklang mit angenommenen physikalischen Wirkungen.

Unsere Bereitschaft, in Bewegung ursächliche Wirkzusammenhänge zu sehen, ist also außerordentlich hoch. Wir sind sogar bereit Zusammenhänge herzustellen, wo räumliche und zeitliche Kontiguität weitgehend verletzt sind. Die Neigung, fast beliebige Ereignisse kausal zu verknüpfen, wenn nur eine gewisse zeitliche Verbindung zwischen ihnen besteht, hat sich das klassische Konditioniern zunutze gemacht. Das Argument, daß wir im Laufe der Evolution so strukturiert wurden, in allem bedeutungsvolle Wirkzusammenhänge zu erleben, hat Riedl (1995) eindrucksvoll dargelegt. Er zeigt auch auf, daß diese Strategie besonders erfolgreich ist, um Wirkzusammenhänge, die tatsächlich bestehen, im Laufe der Evolution auch aufzuspüren. Es ist also ökonomischer, Ursachen auch dort zu vermuten, wo keine bzw. andere bestehen, als ganz darauf zu verzichten.

**Intuitive Physik als Ergebnis eines notwendigerweise unscharfen Mechanismus Wirkursachen zu erkennen**

Unsere Anlage, ja das Bedürfnis, Ursachenbeziehungen wann immer möglich herzustellen, hat dazu geführt, daß wahre Wirkzusammenhänge nur in soweit erkannt bzw.

vom Wahrnehmungssystem internalisiert worden sind, wie dies für das Fortbestehen der Spezies notwendig war. Wo falsch wahrgenommene Wirkzusammenhänge folgenlos blieben, wurden sie nicht korrigiert. So haben wir etwa i. d. R. ein visuelles und auch intuitives Verständnis vieler Vorgänge der klassischen Mechanik, das eine - wenn auch sehr grobe - Annäherung an die Newton'sche Mechanik darstellt. Oft wurde argumentiert, daß sich unser Wahrnehmungssystem etwa auf dem Stand Aristotelischer Physik befindet (z. B. McCloskey, Caramazza & Green, 1980). Ein Beispiel hierfür ist die Überzeugung, daß eine Murmel, die man durch ein gebogenes Schlauchstück rollt, auch nach Austritt auf einem gekrümmten Pfad (B) weiterrollt (siehe Abbildung 3). A ist natürlich Newtons Lösung. Nun hat interessanterweise das visuelle System einen engeren Toleranzbereich als es unser bewußtes Nachdenken über solche Sachverhalte hat. Rein visuell dargeboten ist unser Wissen oft noch näher an der Wahrheit (Kaiser, Proffitt, Whelan & Hecht, 1992), d. h. Pfad B sieht komisch aus, wenn solch ein Ereignis simuliert wird, und zwar auch für Versuchspersonen, die fest von B als richtiger Antwort überzeugt waren.

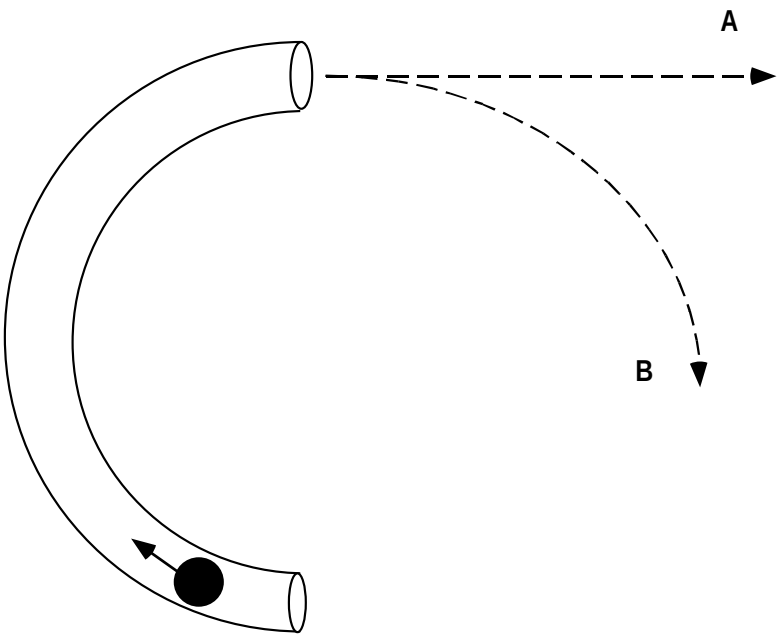



Abbildung 3

**Die Unzulänglichkeit dieser Sichtweise**

Bisher ließe sich der Stand der Dinge so zusammenfassen, daß unsere visuelle Wahrnehmung als Näherungslösung tatsächliche Wirkzusammenhänge erfaßt. Allerdings gibt es schlagende Beispiele von einfachen Situationen, wo diese Behauptung in sich zusammenbricht: Im Falle von ballistischen Bewegungen gehen erschreckend viele Beobachter davon aus, daß sich ein Projektil nach Verlassen der Kanone, aus der es

abgefeuert wird, oder ein Ball nach Verlassen des Werfers Hand, noch beschleunigt. Gleichermaßen wird ein geworfener Ball, der im Steigflug seine Geschwindigkeit noch erhöht als natürlicher beurteilt als Fälle in denen er physikalisch richtig sich verlangsamt (Hecht & Bertamini, im Druck). Abbildung 4 zeigt die Medianwerte der angenommenen maximalen Geschwindigkeit von zwei Ballwürfen, die von 84 durchschnittlichen Collegestudenten erhoben wurden (das 95% Konfidenzintervall ist grau unterlegt, die Box begrenzt erstes und drittes Quartil). Dieser überraschende aber stabile Befund läßt sich nur deuten, wenn wir unsere Körpermotorik - die Beschleunigung des Wurfarms - gleichsam auf das Wurfgeschoß projizieren. Wir dehnen also das, was für unsere Körpermotorik zutrifft, auf die von uns bewegten Dinge aus. Man könnte hier von einer Motortheorie der wahrgenommenen Kausalität sprechen.

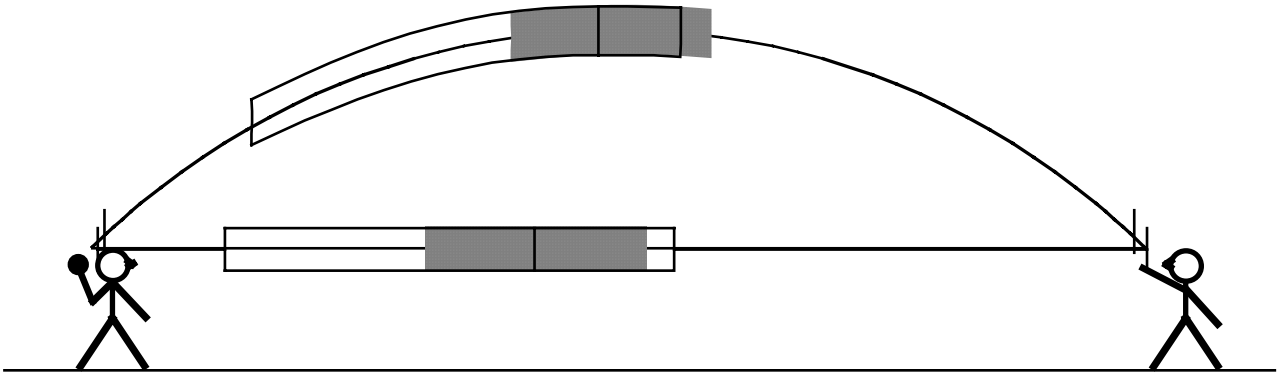


Abbildung 4

**Wahrnehmung von Kausalbeziehungen ist Wahrnehmung von Zweckursachen**

Ebensogut können solche Befunde als Beleg dafür gewertet werden, daß die scharfe zeitliche Grenze zwischen Ursache (Armbewegung) und Wirkung (Ballflug) in unserer Wahrnehmung verwischt wird. Oder anders formuliert, Handlung und Wahrnehmung sind zwar kognitive aber keine visuellen Kategorien. Solch ein Mangel an Trennschärfe ist mit der Annahme einer gemeinsamen Kodierung von Wahrnehmung und Handlung vereinbar (siehe Prinz, 1997). Diese Interpretation ist auf den ersten Blick radikal, sie gewinnt aber an Plausibilität, wenn wir berücksichtigen, daß der Hume'sche Ursachenbegriff sehr einseitig ist. Legt man diesen an, begehen wir Wahrnehmungsfehler. Legt man hingegen den Aristotelischen Ursachenbegriff an, so gilt es, in Wirkursache und Zweckursache zu unterscheiden. Im Sinne einer Zweckursache sind Intentionen die Ursache von Bewegung. Unser Wille, den Ball zu werfen, erklärt also dessen Flug. So gesehen können unsere Wahrnehmungsfehler als Hervortreten der Zweck- vor der Wirkursache interpretiert werden.

Vergegenwärtigen wir uns jetzt, daß eine herausragende Aufgabe unserer Wahrnehmung das Erkennen von sozialen Faktoren beinhaltet (Unterscheiden von Angriff und Flucht, schnelle Bewertung der Absicht des sich auf uns Zubewegenden, etc.), so ist eine Verwechslung von Eigenschaften unbelebter Objekte (Wirkung) mit solchen belebter Subjekte (Zweck) sehr nahe gerückt (vergl. dazu Gelman, Durgin & Kaufman, 1995). So ist auch zu erklären, wieso wir in einfache Bewegungen von geometrischen Objekten spontan Absichten, Emotionen, und soziale Beziehungen hineininterpretieren (Heider & Simmel, 1944). Selbst Kleinkinder neigen dazu, einfachste Bewegungen als intentional zu interpretieren (e. g. Corrigan & Denton, 1996; Premack & Premack, 1997). Daß selbst statische Objekte kausal interpretiert werden und die Geschichte der Einflüsse und Deformationen, die sie erfahren haben, auch visuell erfaßt werden zeigt Leyton (1992).

**Wahrgenommene Kausalität in künstlichen Systemen**

Zusammenfassend ist die menschliche Wahrnehmung von Kausalbeziehungen stark von den eigenen Absichten und Zwecken geprägt. Unsere Körpermotorik legte bisher die Grenzen unserer Zwecke fest. Dieser Zusammenhang von Motorik und wahrgenommener Kausalität ist bisher weitgehend vernachlässigt worden. In künstlichen Systemen, sei es in einer fly-by-wire Kontrolle über ein Verkehrsflugzeug oder in einer Teleoperation, die in einer virtuellen Umgebung ausgeführt wird, wird unsere Wahrnehmung von unseren Effektoren (Hände) künstlich getrennt. Evolutionär ist unsere Kausalitätswahrnehmung auf unmittelbare Weise mit unseren Effektoren verbunden gewesen, wir konnten die Wirkung unserer motorischen Handlungen (Greifen, Werfen, Schlagen, etc.) immer sofort und unvermittelt sehen. Visualisierungen in künstlichen Systemen müssen diese natürliche Grenze unserer Kausalitätswahrnehmung berücksichtigen, indem sie intentional verstehbare Zweckursachen sichtbar machen.

**Literatur**


Barr, H. J. (1964). The epistemology of causality from the point of view of evolutionary biology. Philosophy of Science, 31, 286-288.

Bertamini, M. (1993). Memory for position and dynamic representations. Memory & Cognition, 21, 449-457.

Corrigan, R., & Denton, P. (1996). Causal understanding as a developmental primitive. Developmental Review, 16, 162-202.

Gelman, R., Durgin, F. & Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. In D. Sperber, D. Premack, and A. J. Premack (Eds.), Causal cognition: A multidisciplinary debate (pp. 151-184). Oxford: Clarendon Press.

Hecht, H., & Bertamini, M. (in press). Understanding projectile acceleration. Journal of Experimental Psychology: Human Perception and Performance.

Heffner, J. H. (1981). The causal theory of visual perception: Its scientific basis and epistemological implications. International Philosophical Quarterly, 21, 303-332.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. American Journal of Psychology, 57, 243-259.

Houssiadas, L. (1964). An exploratory study of the perception of causality. The British Journal of Psychology: Monograph Supplements No. XXXVI. London: Cambridge University Press.

Kaiser, M. K., Proffitt, D. R., Whelan, S., & Hecht, H. (1992). Influence of animation on dynamical judgments. Journal of Experimental Psychology: Human Perception and Performance, 18, 669-690.

Leyton, M. (1992). Symmetry, causality, mind. Cambridge, MA: MIT Press.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. Science, 210, 1139-1141.

Michotte, A.(1954). La perception de la causalité. Louvain: Publications Universitaires.

Premack, D., & Premack, A. J. (1997). Infants attribute value+- to the goal-directed actions of self-propelled objects. Journal of Cognitive Neuroscience, 9, 848-856.

Prinz, W. (1997). Perception and action planning. European Journal of Cognitive Psychology, 9, 129-154.

Riedl, R. (1992). Wahrheit und Wahrscheinlichkeit: Biologische Grundlagen des Für-Wahr-Nehmens. Berlin: Parey.

# From Covariation to Causation: A Test of the Assumption of Causal Power

Marc J. Buehner, Department of Psychology, University of Sheffield

Patricia W. Cheng, University of California, Los Angeles

"…what is the nature of that evidence which assures us of any real existence and matter of fact, beyond the present testimony of our senses, or the records of our memory?"

David Hume (1777/1902, p. 26)

David Hume posed a fundamental problem in the acquisition of causal knowledge: our sensory input does not contain explicit information about causal relations. Causal strength must be computed from information available to our senses— such as the presence and absence of candidate causes and effects— by some process of induction. Philosophers, learning theorists, social psychologists and cognitive psychologists have debated how such information is used to infer causality. For candidate causes and effects represented by binary variables with a marked and an unmarked value, a long-standing proposal (e.g., Jenkins & Ward, 1965) is that causal judgments are based on

$$\Delta P = P(e \mid c) - P(e \mid \bar{c}), \qquad (1)$$

where $c$ is a candidate cause, $e$ is the effect in question, $P(e \mid c)$ is the probability of $e$ given the presence of $c$, and $P(e \mid \bar{c})$ is the probability of $e$ given the absence of $c$. $\Delta P$ is a measure of the extent to which $c$ and $e$ *covary*, and is often called the *contingency* or *contrast*. The conditional probabilities are estimated by the relevant relative frequencies. If $\Delta P$ is noticeably positive, $c$ is a *generative* cause, and if it is noticeably negative, $c$ is a *preventive* cause. If $\Delta P$ does not noticeably differ from zero, $c$ is independent of $e$ and is noncausal.

Another attempt to explain the mental leap from covariation to causation also declares heritage to Hume: causal judgment reflects "*no more than the strength of the relevant association between the mental representations of cause and effect, with the principles governing such attributions being those of associative learning*" (Shanks & Dickinson, 1987, p. 230). The most influential theory of associative learning has been the Rescorla-Wagner (1972) model (RWM). Although originally proposed to describe Pavlovian conditioning, several researchers have adopted the RWM to explain human causal reasoning (e.g., Shanks, 1985; Shanks & Dickinson, 1987; Wasserman, Elek, Chatlosh, & Baker, 1993). The core of the RWM is a mechanism of error correction or discrepancy reduction over repeated conditioning trials. Learning proceeds by changing the associative strength between a conditioned stimulus CS (e.g., a flash of light) and an unconditioned stimulus US (e.g., a shock). According to this model,

$$\Delta V_{CS} = \alpha_{CS} \cdot \beta_{US} \cdot (\lambda - \Sigma V), \qquad (2)$$

where $\Delta V_{CS}$ is the change in the strength of an association between a CS and a US after a given trial, and $\alpha_{CS}$ and $\beta_{US}$ are learning rate parameters that respectively represent the salience of the CS (e.g., the brightness of the light) and the US (e.g., the intensity of the shock). $\lambda$ represents the actual outcome of the trial and is typically set to 1.0 if the US is present and to 0.0 otherwise. $\Sigma$V is the sum of the associative strengths of all CSs present, and is thus the expected outcome. Learning is the reduction of discrepancy between the expected and the

actual outcome. When this discrepancy ($\lambda$-$\Sigma$V) approximates zero, learning has reached asymptote: the outcome is fully explained by the CSs. Different assumptions about the learning parameter $\beta$ lead to two different versions of the RWM. The *restricted* RWM operates with $\beta$ held constant across trials on which the US is respectively present and absent ($\beta_{US} = \beta_{\overline{US}}$). The *unrestricted* RWM allows values of $\beta$ to differ between these two types of trials; usually the presence of the US is assumed to be more salient than its absence, yielding $\beta_{US} > \beta_{\overline{US}}$ (cf. e.g., Miller, Barnet, & Grahame, 1995)[1]. This inequality is required for explaining the relative validity of generative candidate causes (Shanks, 1991; cf. Rescorla & Wagner, 1972; Wagner, Logan, Haberlandt, & Price, 1968).

Explaining human causal induction with the RWM reduces causal reasoning to associative learning: the candidate cause *c* is mapped onto the CS, the effect *e* onto the US, and the causal power of *c* is mapped onto *c*'s associative strength. When there is only one varying candidate cause in a context, the restricted RWM asymptotically computes $\Delta P$ as the measure of associative strength (Chapman & Robbins, 1990). The unrestricted RWM, however, does not compute $\Delta P$: with $\beta_{US} > \beta_{\overline{US}}$, the (absolute) magnitudes of the judged causal strengths will be smaller as $P(e \mid \overline{c})$, the *base rate* of *e*, increases for any fixed positive or negative $\Delta P$ (see Wasserman et al., 1993). But if $\beta_{US} < \beta_{\overline{US}}$, contrary to the typical assumption, the RWM would predict the opposite trend: the (absolute) magnitudes of the judged causal strengths should be larger as $P(e \mid \overline{c})$ increases for any fixed positive or negative $\Delta P$. Regardless of assumptions about $\beta$, the RWM predicts $P(e \mid \overline{c})$ to influence the absolute causal strengths for candidates with equal positive $\Delta P$ in the same direction as those with equal negative $\Delta P$.

Kant (1781/1965) and others have noted that covariation does not necessarily imply causation (a point typically emphasized in any course on experimental design or statistics). However, both the contingency model and the RWM disregard Kant's arguments and treat covariation as a direct measure of causation. Neither approach represents causal strength as a variable separate from covariation. As a result, neither can acquire a state of knowledge in which causal strength is unknown (i.e., has no value), and distinguish it from all possible states of knowledge in which causal strength has some value or range of values (Wu & Cheng, 1999). Consider a concrete illustration of this problem. Suppose a researcher wants to evaluate the preventive power of a new headache relieving drug. In a study involving 16 participants, 8 receive treatment with the drug (candidate present), and 8 receive a placebo (candidate absent). Now suppose neither the 8 participants who received treatment nor the 8 control participants report headaches. $\Delta P$ equals 0, yet the researcher would not infer that the drug is ineffective (i.e., noncausal). Because headaches did not occur even in the control group, how then could a preventive candidate prevent them in the treatment group? The researcher simply cannot draw any causal inferences, even though the covariation has a definite value, 0.

To account for this and other problems associated with covariation-based approaches, Cheng (1997) proposed the power PC theory, an integration of covariation-based approaches to causality (e.g., Jenkins & Ward, 1965) and power views (e.g. Michotte, 1946/1963; Bullock, Gelman, & Baillergeon, 1982). According to this theory, the causal reasoner's goal is to induce the unobservable causal power of a candidate cause from observable events. The theory shows that when causes alternative to the candidate cause *c* both occur and influence *e* independently of *c*, and $\Delta P$ is nonnegative, the generative power of *c* to produce *e* is

$$q_c = \frac{\Delta P_c}{1 - P(e \mid \bar{c})}. \tag{3}$$

For a nonpositive $\Delta P$, the same set of assumptions except that $c$ may now potentially *prevent* instead of produce $e$ yields

$$p_c = \frac{-\Delta P}{P(e \mid \bar{c})} \tag{4}$$

as the causal power of $c$ to prevent $e$. Whereas the traditional contingency model (Equation 1) and the restricted RWM each treat $\Delta P$ as the sole indicator of causality, the power PC theory implies that covariation is a necessary but insufficient component for the computation of causal power. Under conditions in which Equations 3 and 4 apply, causal power depends not only on $\Delta P$ but also on $P(e \mid \bar{c})$. It follows from these equations that equal levels of $\Delta P$ with different values of $P(e \mid \bar{c})$ should yield different causal judgments. When evaluating generative causal power, candidate causes with the same positive $\Delta P$ should be judged to have increasingly *large* generative power as $P(e \mid \bar{c})$ increases, but does not equal 1. When $P(e \mid \bar{c}) = 1$, $q_c$ is undefined according to Equation 3: a reasoner cannot draw any conclusion about the causal strength of $c$ generating $e$ if $e$ occurs all the time regardless of $c$. In contrast, candidate causes with equal negative values of $\Delta P$ should be judged to have increasingly *small* preventive power as $P(e \mid \bar{c})$ increases towards 1. When $P(e \mid \bar{c}) = 0$, $p_c$ is undefined according to Equation 4, as in the headache drug example. Finally, Equations 3 and 4 both predict that when $\Delta P = 0$, the power of $c$ should remain at 0 and be uninfluenced by $P(e \mid \bar{c})$ as long as the denominator in the relevant equation is not 0. These predictions are assumed to apply only as ordinal descriptions. Note that (1) $P(e \mid \bar{c})$ influences the (absolute) magnitude of estimated generative and preventive powers in opposite directions, and (2) the direction of these influences is not dependent on any parameter settings. These predictions distinguish the power PC theory from both the RWM and the contingency theory. To illustrate the predictions of the power PC theory, let us return to the researcher who wants to evaluate the effectiveness of new headache relieving drugs: again, 8 participants receive treatment, and 8 a placebo. Now, all 8 participants in the control group have headaches, whereas only 6 of the 8 participants who received the drug have headaches; $\Delta P = P(e \mid c) - P(e \mid \bar{c}) = .75 - 1.00 = -.25$. Assuming that all alternative causes of headaches are constant across the two groups, the researcher will assume that if not for the drug, all 8 participants in the drug group would have had headaches, just as in the control group. The drug therefore has a small preventive power, preventing headaches with a probability of .25. In yet another study, 4 of the 8 participants in the control group report headaches and 2 of the 8 participants in the treatment group report headaches; $\Delta P = P(e \mid c) - P(e \mid \bar{c}) = .25 - .50 = -.25$. Again, assuming that alternative causes remain constant between groups, these causes would have produced headaches in 4 of the 8 participants in the drug group, just as in the control group. The drug therefore prevents headaches in 2 of these 4 participants, yielding a probability of .50. Thus, although $\Delta P = -.25$ here as in the preceding study, the researcher will attribute a higher preventive power to the latter candidate. Equation 4 formalizes this intuition. Analogous intuitions about the generative power of candidate causes with the same positive $\Delta P$ but different $P(e \mid \bar{c})$ are captured by Equation 3.

## Experimental Tests of the Three Models

We tested candidates with positive, negative, and zero $\Delta P$s to evaluate the full range of predictions of Equations 3 and 4 and to cover all possible consistent parameter settings of the RWM. The contingency model and the restricted RWM both predict causal judgments to be solely determined by $\Delta P$. The unrestricted RWM allows causal judgments of candidates with the same $\Delta P$ to be a function of $P(e \mid \bar{c})$, but unless the ordering of its parameter settings is reversed across conditions, it cannot predict a positive function in one and a negative function in another experimental condition. In contrast, the power PC theory predicts opposite functions for generative and preventive conditions. The distinct functions are based solely on the sign of $\Delta P$, which can be computed from the observed input, and are not dependent on any parameter settings.

A test of these three accounts requires testing combinations of many levels of $P(e \mid c)$ and $P(e \mid \bar{c})$. Wasserman et al. (1993) reported the most comprehensive study of this nature to date, but used an effect that occurred at a rate (frequency per unit of continuous time), rather than in a proportion of discrete entities. The power PC theory is a probabilistic theory, and hence does not apply to the type of effect in Wasserman et al. We therefore adapted their design in which 5 levels of $P(e \mid c)$ and of $P(e \mid \bar{c})$ are independently combined within-subject, and modified it by using a binary effect that occurred in discrete entities instead. Because such entities required a longer presentation time, we presented two separate groups of participants with conditions involving nonnegative $\Delta P$s and those involving nonpositive $\Delta P$s to reduce the number of conditions per participant.

## Method

*Participants, design, and procedure.* Fifty-seven (preventive component) and 52 (generative component) undergraduate psychology students from the University of California, Los Angeles participated in exchange for course credit. Participants in the preventive part (involving nonpositive $\Delta P$s) were given a cover story asking them to pretend they were virologists testing several new vaccines against viruses. Each participant worked on one practice condition and 15 experimental conditions, with each condition consisting of 16 laboratory records (i.e., learning trials). Each record provided information about whether or not one particular rat was vaccinated prior to virus exposure, and whether or not this rat developed the disease related to the virus. The generative component (involving nonnegative $\Delta P$s) adopted the same design and procedure, except that participants were asked to imagine that they were microbiologists studying how ray exposure influences the mutation of viruses, and laboratory records informed participants whether or not one particular petri dish with viruses was exposed to certain rays, and whether or not the viruses in this dish mutated. The 15 experimental conditions in both components of the experiment represented 15 independent studies on different viruses and vaccines, or rays and viruses. We used 16 fictitious viruses, vaccinations, and rays.

For the evaluation of preventive causal power, 5 levels of the conditional probabilities $P(e \mid c)$ and $P(e \mid \bar{c})$—1.00, .75, .50, .25, and .00— were combined to yield 5 levels of nonpositive $\Delta P$s: −1.00, −.75, −.50, −.25, and .00. These combinations rendered a total of 15 conditions (see Table 2) which were presented in random order. For the evaluation of generative causal power, the design was exactly symmetrical, producing nonnegative $\Delta P$s with values .00, .25, .50, .75, and 1.00 (see Table 1).

The 16 laboratory records for the evaluation of each candidate consisted of 8 for which *c* was present (the rat was vaccinated, or the petri dish was exposed to rays) and 8 for which *c* was absent (no vaccination given, or no ray exposure). These records were presented

in random order sequentially on a computer screen. For each record, participants had to confirm the information provided by clicking appropriate checkboxes on the computer screen ('received vaccination? yes/no' and 'broke out with virus disease? yes/no' in the preventive part; and 'was exposed to rays? yes/no' and 'mutated? yes/no', in the generative part).

After having studied the 16 records in a condition, participants had to evaluate the causal power of the studied candidate. Participants in the preventive part were asked to judge how strongly each vaccine prevented the disease related to the virus in question by giving a rating on a scale from 0 (the vaccine *does not prevent the disease at all*) to 100 (the vaccine prevents the disease *every time*). Analogously, participants in the generative part of the study were asked to rate how strongly they thought the particular rays cause mutation on a scale from 0 (the ray *does not cause mutation at all*) to 100 (the rays *cause mutation every time*). To keep the judgments simple, the task did not provide the option of answering "I don't know". Following their answer, participants' confidence in their judgment as well as their subjective probability ratings of $P(e \mid c)$ and $P(e \mid \bar{c})$ were collected.[2] The questions were always presented sequentially in the order just described. A graphical scale with corresponding labels at the endpoints and numerical markings at regular intervals was presented along with the text for the causal and confidence ratings. All participants were run individually, and the experiment took approximately 45 minutes.

## Results and Discussion

Figure 1 displays participants' mean ratings of the preventive and generative causal power of the candidate causes. On the abscissa are the 5 levels of $P(e \mid \bar{c})$. The lines connect mean ratings for the same level of $\Delta P$. (Corresponding values for $P(e \mid c)$ follow from each combination of $\Delta P$ and $P(e \mid \bar{c})$. The means and standard deviations are listed in Tables 1 and 2.)

A visual examination of Figure 1 reveals a substantial influence of $P(e \mid \bar{c})$ on causal judgments for candidates with the same $\Delta P$. As predicted uniquely by the power PC theory, this influence occurred in opposite directions for the preventive and generative causal ratings. We used one-factor repeated–measures ANOVAs to examine trends due to $P(e \mid \bar{c})$ in conditions with equal levels of $\Delta P$, employing an alpha level of .05 in all statistical tests. Linear negative trends were highly reliable for each level of negative $\Delta P$: $t(56)=4.14$, $p<.001$, for $\Delta P= -.75$; $t(112)=9.48$, $p<.001$, for $\Delta P=-.50$; and $t(168)=13.34$, $p<.001$ for $\Delta P=-.25$, and no quadratic or cubic trends were found. For the generative part of the study, causal judgments for equal levels of $\Delta P$ generally increased as $P(e \mid \bar{c})$ increased. For $\Delta P$ =.75, however, the positive linear trend fell short of significance, $t(51)=1.99$, $p=0.52$. Linear positive trends were reliable for $\Delta P=.50$, $t(102)=3.72$, $p<.001$ (in addition to a cubic trend, $t(102)=3.26$, $p<.002$), and for $\Delta P=.25$, $t(153)=3.91$, $p<.001$.

This pattern of negative trends for candidates with the same negative $\Delta P$ and positive trends for candidates with the same positive $\Delta P$ clearly contradicts the traditional contingency model and the restricted RWM, as both postulate $\Delta P$ as the sole determinant of causal judgments and thus predict flat lines. The results also contradict the unrestricted version of the RWM with a consistent ordering of parameter settings, which could explain trends for equal levels of $\Delta P$, but merely in one single direction across preventive and generative scenarios. Only the power PC theory correctly predicts the pattern of opposite trends for generative and preventive scenarios.

Participants' evaluations of non-contingent candidate causes substantially deviated from zero. This result replicates earlier observations (e.g., Shanks, 1985, 1987), but is not predicted by any of the accounts considered here, all of which predict that zero $\Delta P$s should

yield non-causal ratings. Inspection of our data suggests a base-rate influence on causal ratings for non-contingent candidates. In the preventive study, the 4 non-contingent candidates with defined causal powers (see Table 2) produced a negative linear trend, $t(168)=9.57$, $p<.001$, and the 4 corresponding non-contingent candidates in the generative study (see Table 1) produced a positive linear trend, $t(153)=7.59$, $p<.001$, and a cubic trend, $t(153)=3.52$, $p<.001$. Because participants in our studies gave ratings on a scale from 0 to 100, judging causal power to be 0 requires giving ratings at an extreme end of the scale. Deviations from normative values at extreme ends of a scale can only occur in one direction, away from the extreme, potentially producing artificially elevated data.[3] This artifact, however, cannot by itself explain the observed trends as a function of the base-rate.

In addition to the scaling problem, participants in our experiments could have conflated reliability with causal strength. For example, in the generative "0/8-0/8" condition ($P(e \mid c) = P(e \mid \bar{c}) = 0$), *e* occurred neither in any of the 8 trials given the presence of *c* nor in those given its absence. Assuming that alternative causes were constant, it follows that *c* had 8 trials to prove its generative power, and it failed to do so. In the 6/8-6/8 condition ($P(e \mid c) = P(e \mid \bar{c}) = .75$), however, *e* would be expected to be produced by alternative causes on 6 of the 8 trials when *c* was present, just as when *c* was absent. This leaves only 2 out of the 8 trials for *c* to prove its causal power, and it failed to do so. Providing a constant number of trials across conditions therefore leads to varying reliability between these conditions: because there were more trials on which *c* could have but in fact failed to prove its power in the former condition than in the latter, participants might be more confident of a noncausal rating in the former condition, leading to a causal rating closer to 0. Analogously, in the 8/8-8/8 condition of the preventive study, *e* occurred on all of the 8 trials where *c* was absent, creating 8 chances for *c* to prove its preventive power where it failed to do so. On the other hand, *c* only had 2 trials to show its preventive power in the 2/8-2/8 condition.

It is important to emphasize that this reliability explanation does not apply to conditions with equal levels of non-zero $\Delta P$s. For example, for preventive candidates that all have $\Delta P = -.5$, in the 0/8–4/8 condition, $p_c = 1.0$; in the 2/8–6/8 condition, $p_c = .67$; and in the 4/8–8/8 condition, $p_c = .5$. If one were to ignore the varying causal powers, and consider reliability alone, the three conditions just mentioned should have increasingly *greater* reliability: The 0/8–4/8 condition had only 4 trials to reveal preventive power, the 2/8–6/8 condition had 6 trials, and the 4/8–8/8 condition had 8 trials. Therefore, given that these candidates have nonzero preventive power, reliability predicts increasing preventive ratings for these three conditions, contrary to the observed decreasing ratings, which were as predicted by their decreasing causal powers. Reliability thus cannot explain the trends observed for candidates with equal non-zero $\Delta P$s.

Our experimental design also allows comparisons between conditions with varying $\Delta P$s for which causal power is predicted to be constant. However, unlike the comparisons between candidates with equal $\Delta P$s but unequal powers, these comparisons are confounded with reliability. Among the preventive conditions, the 0/8–2/8, 0/8–4/8, 0/8–6/8, and the 0/8–8/8 conditions all yield $p_c = 1$. In addition, the 2/8–4/8 and 4/8–8/8 conditions both yield $p_c = .5$. For these conditions, reliability increased as base rate increased: for example, while the 0/8–2/8 condition left only 2 trials in which *c* could and did successfully prove its preventive power, the 0/8–8/8 condition had 8 trials in which *c* could and did prevent *e*. Analogously there are four generative conditions with $q_c=1$ (8/8–0/8, 8/8–2/8, 8/8–4/8, and 8/8–6/8) and two with $q_c=.5$ (4/8–0/8 and 6/8–4/8). For these conditions, reliability decreased as base rate increased: for example, *c* had 8 trials on which it could and did produce *e* in the 8/8–0/8 condition, as compared to only 2 trials in the 8/8–6/8 condition.

We analyzed conditions with equal causal power with repeated measures ANOVAs with $\Delta P$ as the within-subject factor. Linear trends were reliable in the preventive conditions, $t(168)=5.82$, $p<.001$, and the generative conditions, $t(153)=6.16$, $p<.001$, with causal powers of 1. Conditions with power equal to .5 differed significantly, $t(56)=3.54$, $p<.001$ for the preventive candidates, and $t(51)=2.26$, $p<.03$ for the generative candidates. The varying reliabilities for candidates with unequal $\Delta P$s but equal powers can explain every of these trends except one (the 2/8–4/8 condition, which had a lower reliability than the 4/8–8/8 condition, yielded a higher preventive rating, even though in both conditions, $p_c$=.5).

## Summary and Conclusion

Our experiments contrasted the predictions of the power PC theory with the traditional contingency model and the RWM. Our results show that, as predicted uniquely by the parameter-free power PC theory, the base rate of the effect influences causal estimates of candidates with equal non-zero $\Delta P$s, and interacts with the causal scenario in determining the direction of the influence. Whereas the observed deviations from the predictions of this theory can be explained by participants' conflation of reliability and strength, this interaction cannot.

The contingency model and the restricted RWM, which do not allow any influence of base rate, are clearly refuted. The unrestricted RWM, which allows different parameter values for $\beta$, cannot explain our results as long as a consistent set of parameters is applied. It can explain the negative trends observed for the preventive candidates if $\beta_{US} > \beta_{\overline{US}}$, and the positive trends observed for the generative candidates if the parameter setting is reversed ($\beta_{US} < \beta_{\overline{US}}$), but the latter setting would contradict that required to explain relative validity, a well-established phenomenon involving generative candidates (Shanks, 1991; cf. Rescorla & Wagner, 1972; Wagner, Logan, Haberlandt, & Price, 1968). Thus, the RWM—even in its unrestricted form— cannot satisfactorily explain our results.

Lober and Shanks (in press), in a critique of a preliminary report of the data presented here (Buehner & Cheng, 1997), argue that a reversal in the RWM's parameter settings between scenarios is justified and reflects a difference in the salience of *e* across the preventive and generative scenarios. This argument provides only an untested post-hoc explanation of the data reported here, and does not resolve the contradictory requirement of opposite parameter settings to account for results across studies all involving generative candidates.

Our findings show that purely covariation-based models, including the associationist RWM, do not adequately describe human causal induction. They suggest that the mental leap from covariation to causation requires that the reasoner hold the conviction that there are unobservable causal powers in the environment, and that the goal of causal induction is to infer them.

## References


Buehner, M. J., & Cheng, P.W. (1997). Causal induction: The power PC theory versus the Rescorla-Wagner model. In the *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society (pp. 55-60).* Hillsdale, NJ: Erlbaum.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp209-254). New York: Academic Press.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18(5),* 537-545.

Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367-405.

Hume, D. (1902). An enquiry concerning human understanding. In: L.A. Selby-Bigge (Ed.), *Hume's Enquiries.* Oxford, England: Clarendon Press. . (Original work published 1777).

Kant (1965). *Critique of pure reason*. London: Macmillan. (Original work published 1781).

Lober, K., & Shanks, D. R. (in press). Experimental falsification of Cheng's (1997) power PC theory of causal induction. *Psychological Review*.

Michotte, A. E. (1963). *The perception of causality.* New York: Basic Books. (Original work published 1946).

Neunaber, D. J., & Wasserman, E. A. (1986) The effects of unidirectional versus bidirectional rating procedures on college students' judgments of response-outcome contingency. *Learning and Motivation 17*, 162-179.

Miller, R.R., Barnet, R.C., & Grahame, N.J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin 117*, 363-386.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), Classical conditioning II: Current theory and research (pp. 64-99). New York: Appleton-Century Crofts.

Shanks, D. R. (1985). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition, 13*, 158-167.

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation, 18*, 147-166.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 433-443.

Shanks, D., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. Bower (Ed.), *The psychology of learning and motivation, Vol 21* (pp. 229-261). San Diego, CA: Academic Press.

Wagner, A.R., Logan, F.A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology, 76*, 171-180.

Wasserman, E.A., Elek, S.M., Chatlosh, D.L., & Baker, A.G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 174-188.

Wu, M., & Cheng, P.W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science, 10*, 92-97.

## Author Note

This paper is based on a master's thesis submitted by the first author to the Department of Psychology, University of California, Los Angeles, under the supervision of the second author. Part of this work was presented at the 19th Annual Meeting of the Cognitive Science Society in Stanford, CA, 1997, and at the 39th Tagung experimentell arbeitender Psychologen in Berlin, Germany, 1997. Preparation of this article was supported by a Fulbright scholarship and a Friedrich-Ebert grant awarded to the first author, and by NSF Grant 9729726 to the second author. Final completion of the manuscript was supported by the EU Tranining and Mobility of Researchers network TACIT. Sean McAulliffe programmed the experimental materials. We thank Merideth Gattis and Keith Holyoak for valuable comments on an earlier draft. Correspondence concerning this

article should be sent to Marc Buehner, Department of Psychology, University of Sheffield, Sheffield, S10 2TP, England, UK. Electronic mail may be sent via the Internet to marc.buehner@sheffield.ac.uk.

## Footnotes

1 Lober and Shanks (in press) coined the terms *restricted* versus *unrestricted* Rescorla-Wagner model.

2 We will only report the causal ratings. Confidence ratings mirrored the patterns found in causal ratings and subjective probability ratings are not important for the purpose of this paper.

3 We thank Tom Wickens for pointing out this restriction in the interpretation of scores obtained at the extreme ends of rating scales. Corroborating the case for artifacts of measurement, Neunaber and Wasserman (1986) reported that zero-contingencies were rated less accurately when participants had to use a unidirectional scale from 0 to 100 compared to a bidirectional scale from -100 to 100.

Table 1. Base rate effect on causal judgments for equal levels of $\Delta P$, generative scenario.

Generative Scenario

| $P(e \mid c)$ | $P(e \mid \bar{c})$ | ΔP | Causal power | Causal ratings | |
|---|---|---|---|---|---|
| | | | | Means | Std Dev |
| 1.00 | 0.00 | 1.00 | 1.00 | 88.54 | 20.05 |
| | | | | | |
| 1.00 | 0.25 | 0.75 | 1.00 | 76.54 | 23.86 |
| 0.75 | 0.00 | 0.75 | 0.75 | 69.02 | 22.21 |
| | | | | | |
| 1.00 | 0.50 | 0.50 | 1.00 | 70.85 | 23.19 |
| 0.75 | 0.25 | 0.50 | 0.67 | 54.02 | 23.77 |
| 0.50 | 0.00 | 0.50 | 0.50 | 57.00 | 22.73 |
| | | | | | |
| 1.00 | 0.75 | 0.25 | 1.00 | 57.62 | 32.35 |
| 0.75 | 0.50 | 0.25 | 0.50 | 47.06 | 26.49 |
| 0.50 | 0.25 | 0.25 | 0.33 | 45.71 | 22.53 |
| 0.25 | 0.00 | 0.25 | 0.25 | 33.69 | 26.67 |
| | | | | | |
| 1.00 | 1.00 | 0.00 | undefined | 41.10 | 36.85 |
| 0.75 | 0.75 | 0.00 | 0.00 | 42.62 | 27.06 |
| 0.50 | 0.50 | 0.00 | 0.00 | 36.60 | 24.91 |
| 0.25 | 0.25 | 0.00 | 0.00 | 28.62 | 21.45 |
| 0.00 | 0.00 | 0.00 | 0.00 | 9.46 | 20.03 |

Table 2. Base rate effect on causal judgments for equal levels of $\Delta P$, preventive scenario.

Preventive Scenario

| $P(e \mid c)$ | $P(e \mid \bar{c})$ | ΔP | Causal power | Causal ratings | |
|---|---|---|---|---|---|
| | | | | Means | Std Dev |
| 0.00 | 1.00 | -1.00 | 1.00 | 93.58 | 12.87 |
| 0.00 | 0.75 | -0.75 | 1.00 | 84.75 | 19.72 |
| 0.25 | 1.00 | -0.75 | 0.75 | 71.54 | 17.26 |
| 0.00 | 0.50 | -0.50 | 1.00 | 79.28 | 20.54 |
| 0.25 | 0.75 | -0.50 | 0.67 | 64.61 | 19.49 |
| 0.50 | 1.00 | -0.50 | 0.50 | 45.95 | 21.87 |
| 0.00 | 0.25 | -0.25 | 1.00 | 72.00 | 29.62 |
| 0.25 | 0.50 | -0.25 | 0.50 | 58.56 | 20.02 |
| 0.50 | 0.75 | -0.25 | 0.33 | 42.86 | 21.83 |
| 0.75 | 1.00 | -0.25 | 0.25 | 21.09 | 15.8 |
| 0.00 | 0.00 | 0.00 | undefined | 45.12 | 41.05 |
| 0.25 | 0.25 | 0.00 | 0.00 | 47.79 | 28.70 |
| 0.50 | 0.50 | 0.00 | 0.00 | 33.89 | 23.06 |
| 0.75 | 0.75 | 0.00 | 0.00 | 22.86 | 21.86 |
| 1.00 | 1.00 | 0.00 | 0.00 | 8.68 | 20.39 |

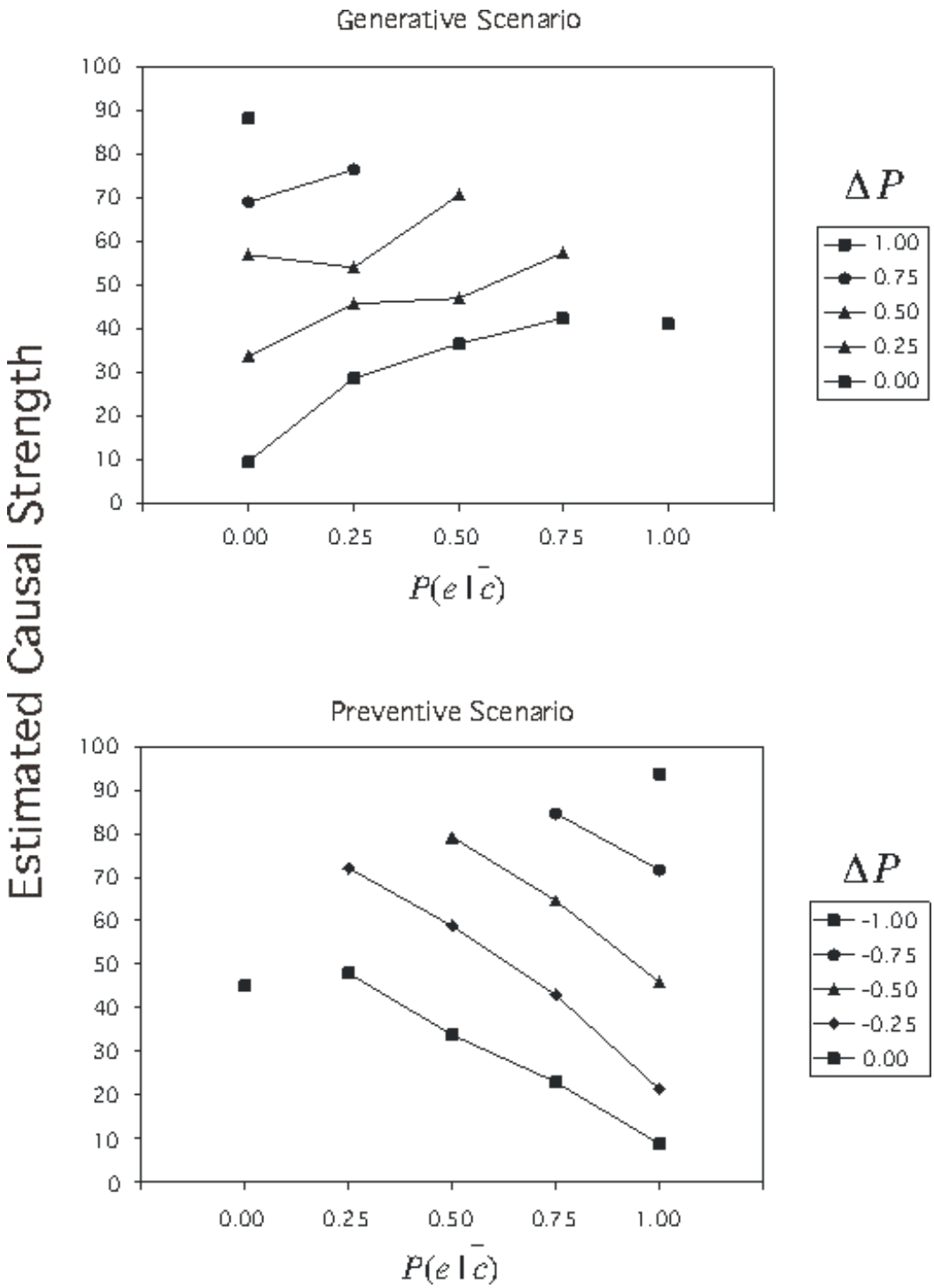



*Figure 1.* Mean estimates of preventive and generative causal powers. Connected lines share the same level of $\Delta P$, but have different causal powers. The unconnected single points with zero $\Delta P$ have undefined causal power.

# Bayessche Netze als Modelle der Repräsentation von Kausalwissen

Marcus Plach
Fachrichtung Psychologie
Universität des Saarlandes
marcus@cops.uni-sb.de

## 1. Einleitung

Die Aktualisierung von Überzeugungsstärken aufgrund neuer Information ist ein fundamentaler Prozeß unseres kognitiven Systems und Voraussetzung für situationsangemessenes Verhalten. Bei der kognitionspsychologischen Untersuchung von Prozessen der Aktualisierung von Überzeugungsstärken (*belief updating*) wurde in der Vergangenheit die Bedeutung von vorhandenem Wissen für die Evaluation neuer Information vernachlässigt. In der vorliegenden Arbeit wird argumentiert, daß probabilistische Kausalmodelle die *constraints* liefern, die ein rationales *belief updating* ermöglichen, und daß Menschen für die Implikationen verschiedener Ursache-Wirkungs-Modelle für das Schlußfolgern unter Unsicherheit sensitiv sind. Neben Varianten probabilistischer Kausalitätstheorien (z.B. Eells, 1991; Cartwright, 1989), die sich beispielsweise in der Theorie der Kausalmodelle niedergeschlagen haben (Waldmann & Holyoak, 1992; Waldmann, 1996), stellte in den letzten Jahren vor allem die Theorie Bayesscher Netze (z.B. Pearl, 1988) einen Bezugspunkt für die kognitionspsychologische Modellbildung zum kausalen Lernen und Schlußfolgern dar. Im folgenden wird untersucht, in welcher Hinsicht Bayessche Netze (BN) ein geeignetes Rahmenmodell für die kognitive Repräsentation von Annahmen über probabilistische Kausalzusammenhänge und der darauf operierenden Inferenzprozesse darstellen. Es wird außerdem beleuchtet, inwieweit empirische Befunde zum Kausallernen mit der Theorie der Bayesschen Netze verträglich sind. Schließlich wird eine experimentelle Untersuchung beschrieben, die zeigt, daß BN wesentliche Eigenschaften der kognitiven Repräsentation und Verarbeitung von Kausalwissen abbilden.

## 2. Die grundlegende Semantik Bayesscher Netze

### *2.1 Wissensrepräsentation*

Bei Bayesschen Netzen handelt es sich um gerichtete azyklische Graphen. Ein Knoten steht für eine Partition unsicherer propositionaler Ausdrücke, d.h. er repräsentiert eine diskrete Zufallsvariable. Eine Variable „Lungenkrebs“ könnte beispielsweise die Propositionen „Die Person wird an Lungenkrebs erkranken“ und „Die Person wird nicht an Lungenkrebs erkranken“ ausdrücken. Überzeugungsgrade korrespondieren in einem BN mit einer Wahrscheinlichkeitsverteilung über die Werte der durch die Knoten repräsentierten diskreten Zufallsvariablen. Die Kanten eines Netzes repräsentieren Abhängigkeiten zwischen den Variablen. Die Stärke dieser Abhängigkeiten wird in einem BN durch bedingte Wahrscheinlichkeiten quantifiziert. Ein BN repräsentiert die *gemeinsame Wahrscheinlichkeitsverteilung* der Variablen, die in einer Domäne im Rahmen eines Wahrscheinlichkeitsmodells betrachtet werden: $P(x_1,...,x_n) = \prod_{i=1}^{n} P(x_i \mid Eltern(X_i))$, wobei die Menge *Eltern* diejenigen Knoten enthält, von denen aus direkte Kanten zum Knoten $X_i$ zeigen.

Ein BN kann aber grundsätzlich auch als Enkodierung einer Sammlung von Annahmen bedingter Unabhängigkeiten in der betrachteten Domäne aufgefaßt werden (s. Russell & Norvig, 1995). Der Grund liegt darin, daß ein BN als sogenanntes minimales *I-Map* definiert ist, d.h. die im Graphen als bedingt unabhängig angezeigten Knotenmengen repräsentieren im Wahrscheinlichkeitsmodell bedingt unabhängige Variablenmengen. Diese Knotenmengen sind durch das sogenannte *d-Separationskriterium* von anderen unterscheidbar (vgl. Pearl, 1988; Plach, 1998). Umgekehrt sind alle Abhängigkeiten im Wahrscheinlichkeitsmodell auch im BN durch Knotenmengen repräsentiert, die das *d-Separationskriterium* nicht erfüllen. Die in einem BN repräsentierten bedingten Unabhängigkeitsannahmen sind zentral für die Interpretation von BN als Kausalmodelle, denn bestimmte Zusammenhangsstrukturen von Variablen und den zwischen ihnen existierenden bedingten Unabhängigkeiten reflektieren elementare Kausalmodelle (s. Abschnitt 4).

### *2.2 Updating*

Die Logik der Informationsverarbeitung in einem BN läßt sich als eine Anwendung eines verallgemeinerten Bayes Theorems für den multivariaten Fall beschreiben. *Updatin*g ist ein zwei-stufiger Prozeß. Der Prozeß wird angestoßen durch die Instantiierung eines oder mehrerer Wurzel- oder Blattknoten (d.h. Knoten, die keine Vorgänger- bzw. Nachfolgerknoten aufweisen). Im nächsten Schritt erfolgen dann in iterativen Zyklen lokale Berechnungen an einzelnen Knoten und die Ausbreitung der Änderungen über das gesamte Netz (für technische Details siehe z.B. Pearl, 1988; Neapolitan, 1990). Die in einem BN repräsentierten Unabhängigkeitsannahmen sind auch die Voraussetzung dafür, daß jede Subkomponente auf direktem Wege nur mit einer begrenzten Zahl von anderen Subkomponenten interagiert, unabhängig von der gesamten Zahl der Komponenten (BN werden deshalb auch als *lokal-strukturierte* oder *modular-strukturierte* Systeme bezeichnet). Dieser Sachverhalt wirkt sich auch auf die Zahl der zu repräsentierenden Wahrscheinlichkeiten aus (s. Abschnitt 2.3). Der *Updating*-Vorgang ist strukturell nach Spohn (1988) insofern analog zum menschlichen *Updating*-Prozeß als dieser mit der Wahrnehmung eines geänderten oder neuen Sachverhalts über die Sinne beginnt und in einer nicht weiter analysierbaren Weise zu einer subjektiven Wahrscheinlichkeit 1 für die Basis der Änderung führt; diese Änderung an der Basis bewirkt dann gemäß einer verallgemeinerten Konditionalisierungsregel die Änderung des übrigen epistemischen Zustands (vgl. Spohn, 1993).

### *2.3 Komplexität*

Wie bereits unter Abschnitt 2.1 kurz dargestellt, repräsentiert ein BN die *gemeinsame Wahrscheinlichkeitsverteilung* der Variablen in einem Wahrscheinlichkeitsmodell. *Belief updating* und - darauf gestützt - diagnostisches und prädiktives Schlußfolgern unter Unsicherheit, kann in einem wahrscheinlichkeitstheoretischen Kontext als Berechnung von bedingten Wahrscheinlichkeiten aufgefaßt werden (dies ist durch die Konditionalisierungsregel gerechtfertigt, s. Plach, 1998, S. 92f). Es stellt sich die Frage, wie diese Berechnung in einem System am effizientesten durchgeführt werden kann. Geht man von den Verbundwahrscheinlichkeiten der Zustände in der Domäne als Ausgangspunkt aus, dann ließen sich aus Quotienten von Randwahrscheinlichkeiten, die wiederum auf Summen von Verbundwahrscheinlichkeiten zurückgeführt werden könnten, bedingte Wahrscheinlichkeiten berechnen. Dieser Weg würde jedoch das Speichern von exponentiell vielen Wahrscheinlichkeiten erfordern. Im Falle von $n$ binären Variablen ergäben sich z.B. $2^n$

zu speichernde Wahrscheinlichkeiten. Unabhängigkeiten können jedoch Abhilfe schaffen. Würde beispielsweise die Domäne *n* stochastisch unabhängige Variablen aufweisen, dann wären nur *n* Wahrscheinlichkeiten zu speichern und die Verbundwahrscheinlichkeiten könnten als Produkt berechnet werden. Dieser Fall ist jedoch für praktisch jede interessante Domäne ausgeschlossen. In einem BN wird nun ausgenutzt, daß in einer Domäne i.d.R. jedoch Variablen vorkommen, die von anderen unabhängig sind, gegeben eine bestimmte Menge von Variablen, die diese von dem direkten Einfluß anderer Variablen abschirmen. In der Regel können wir in einer Domäne die Anzahl *k* derjenigen Variablen, die eine andere Variable direkt beeinflußen, als deutlich geringer annehmen als *n*. Nehmen wir ein konstantes *k* an, dann benötigen wir nur $n2^k$ Wahrscheinlichkeiten (vgl. auch Russell & Norvig, 1995). Dies bedeutet beispielsweise ganz konkret für eine Domäne sinnvoller Größe: bei $n$=10 wären ohne die Annahme bedingter Unabhängigkeiten 1024, bei $k$=3 dagegen nur 80 Wahrscheinlichkeiten zu speichern.

## 3. Bayessche Netze und Kausalität

Die Bedeutung bedingter Unabhängigkeit für die Modellierung von Kausalwissen und Prozessen der Wissensaktualisierung wird deutlich, wenn man dazu übergeht, bedingte Unabhängigkeit nicht allein als statistische Eigenschaft von Variablen, sondern als eine konstruktive Komponente des Schlußfolgerns unter Unsicherheit zu betrachtet:

*Contrary to positions often found in the literature, conditional independence is not a restrictive assumption made for mathematical convenience; nor is it an occasional grace of nature for which we must passively wait. Rather it is a mental construct that we actively create, a psychological necessity that our culture labors to satisfy (Pearl, 1988, S. 385).*

Ganz ähnlich wird die Bedeutung von bedingten Unabhängigkeitsannahmen in der Theorie der Kausalmodelle gesehen. Das in Abschnitt 2.3 dargestellte Zahlenbeispiel verdeutlicht, wie durch die Nutzung bedingter Unabhängigkeit die zu erfassende Information eines Lernsystems reduziert werden kann. Das sich in bedingten Unabhängigkeiten widerspiegelnde Strukturwissen erlaubt es dem Lernsystem also, die zu erfassenden Information zu minimieren, und kommt damit unserer begrenzten Informationsverarbeitung entgegen (Waldmann & Hagmayer, 1998, S. 106). Diese Theorie nimmt des weiteren an, daß die Akquisition von kausalem Wissen durch eine Interaktion von daten- und wissensgetriebenen Prozessen charakterisierbar ist. Es ist zwar bekannt, daß domänen-spezifisches Wissen Lernen leitet, die Theorie der Kausalmodelle jedoch geht noch weiter und nimmt an, daß Vorannahmen in Form kausaler Modelle den Lernprozeß leiten, auch wenn keine domänen-spezifische Information gegeben ist. Darüber hinaus wird angenommen, daß Kausalmodelle richtungsspezifisch und zwar von der Ursache zur Wirkung erworben werden. Beide Inferenzrichtungen sind jedoch für Menschen nach dieser Theorie zugänglich (vgl. Waldmann, 1996).

Die in diesen Formulierungen nahegelegte Interpretation von Bayesschen Netzen als Rahmenmodell für (mentale) Kausalmodelle erfordert allerdings allein schon aus formalen Gründen mindestens drei Zusatzannahmen (vgl. Pearl, 1996): 1. Die temporale Ordnung, 2. die *closed world* Annahme und 3. die technische Annahme des Übergangs von Ereignissen zu Variablen (letztere wird hier nicht weiter erläutert).

*Temporale Ordnung.* Damit ein BN als Kausalmodell von Wirkungsbeziehungen zwischen Ursachen und Effekten aufgefaßt werden kann, muß bei der Generierung des Bayesschen Netzes eine Ordnung der Variablen entlang der Zeitachse vorausgesetzt werden. Für die Konstruktion eines Bayesschen Netzes ist ganz allgemein eine Variablenordnung notwendig. Diese muß nicht notwendig temporal sein. Solange die zentrale Eigenschaft der *Markov-Elternschaft* (s. z.B. Pearl, 1995) eingehalten wird, repräsentieren verschiedene BN exakt die gleiche Verbundwahrscheinlichkeit und sind ein *I-Map*. Diese „Ordnungsbeliebigkeit" ist einer der Hauptgründe, warum Kritiker eine Kausal-Interpretation von BN ablehnen (Schum, 1994). Trotz der Ordnungsbeliebigkeit weisen BN, die nicht auf eine temporale Ordnung zurückzuführen sind, aber im allgemeinen den Nachteil auf, daß sie zu einer größeren Anzahl von Kanten führen, d.h. nicht alle bedingten Unabhängigkeiten in der Domäne werden im Graphen angezeigt. Hinzu kommt, daß bei Quantifizierungen entgegen der temporalen Ordnung Likelihoodschätzungen vorgenommen werden müssen, die für Menschen sehr kontraintuitiv sein können (vgl. Russell & Norvig, 1995). Allgemein gilt: wird eine temporale Ordnung bei der Generierung eines BN zugrundegelegt, sind am Ende weniger Wahrscheinlichkeiten zu spezifizieren, und die Schätzung dieser wird einfacher durchzuführen sein.

*Closed world Annahme. A*lle Variablen in einem BN stellen auch alle *relevanten* Variablen für das Phänomen dar. Hierbei handelt es sich um eine kritische Annahme, mit dem Ziel, auszuschließen, daß Variablen existieren, die im BN vorhandene Variablen aufgrund von *spurious correlations* zu Scheinursachen machen würden. Da derartige potentielle Ursachenvariablen gar nicht erkannt werden, kann ein durch ein BN repräsentiertes Kausalmodell immer nur vorläufigen Charakter haben (allerdings gilt diese Einschränkung für Kausalmodelle jeglicher Provenienz).

## 4. Typen elementarer Kausalmodelle in BN

Werden die in Abschnitt 3 genannten Zusatzannahmen getroffen, so ist es möglich, ein BN als Kausalmodell zu interpretieren. Ein solches Kausalmodell setzt sich dabei aus elementaren kausalen Strukturen zusammen, die im folgenden kurz diskutiert werden sollen.

### *4.1* $B \leftarrow A \rightarrow C$ *-Modelle*

Ein solches Modell impliziert beispielsweise, daß B und C marginal abhängig sind: $P(B,C) \neq P(B)P(C)$, daß aber B bedingt unabhängig von C ist, gegeben A: $P(B \mid A,C) = P(B \mid A)$. So besteht beispielsweise zwischen Körpergröße und Haarlänge (vermutlich auch heute noch) eine negative Korrelation (bei Mitteleuropäern) und damit Abhängigkeit (vgl. Slovic & Lichtenstein, 1971). Kaum jemand wird jedoch annehmen, daß die Körpergröße in einem ursächlichen Zusammenhang mit der Haarlänge steht. Vielmehr würde man annehmen, daß Männer in der Regel größer sind als Frauen und in der Regel per Konvention kürzere Haare tragen als Frauen. Der gemeinsame Grund für das gehäufte Auftreten relativ längerer Haare mit relativ kleinerem Körper (und entsprechend aller anderen Zellen in einer Kontingenztabelle) liegt also in einer Variablen Geschlecht (B). Haarlänge (A) und Körpergröße (C) sind als bedingt unabhängige Effekte einer Ursache Geschlecht (B) aufzufassen. Dies drückt sich aus in einem BN der Struktur $B \leftarrow A \rightarrow C$. Insbesondere drückt die Abwesenheit eines Pfeils zwischen B und C aus, daß $P(B \mid A,C) = P(B \mid A)$. Um eine solche Kausalstruktur entdecken (lernen) zu können, ist es notwendig, *bedingte Kontingenzen* bzw. mit Kontextvariablen konditionierte Wahrscheinlichkeiten zu betrachten. Dies steht in

Einklang mit der Konzeption von konditionalen Kontingenzen wie sie von probabilistischen Kausalitätstheorien als Voraussetzung für die Überprüfung von *spurious correlations* gemacht werden (z.B. Eells, 1991). Danach ist bei potentieller multikausaler Beeinflussung eine Variable C erst dann kausal relevant, wenn alternative kausale Faktoren konstant gehalten werden, und trotzdem gilt: $P(A \mid C, B) > P(A \mid \neg C, B)$.[1]
(Es gilt übrigens: $P(A \mid B, C) = P(A \mid \neg B, C) \Leftrightarrow P(A \mid B, C) = P(A \mid C)$.)
Waldmann und Holyoak (1992) liefern indirekte empirische Evidenz dafür, daß Menschen sensitiv für die Implikationen eines solchen Kausalmodells sind.

### *4.2* $A \rightarrow B \rightarrow C$ *-Modelle*

Eine derartige Struktur impliziert unter anderem, daß zwar einerseits A und C abhängig sein können: $P(A, C) = P(A)P(C)$, andererseits aber auch C bedingt unabhängig ist von A, gegeben B. Zur Erläuterung nehmen wir an, daß eine Person nach kurzem Torkeln auf die Straße läuft und überfahren wird. Die proximale Ursache für den Unfall ist eine Kreislaufschwäche, die zu einer akuten Orientierungslosigkeit und Ohnmachtsnähe geführt hat. Die Ursache der Kreislaufschwäche ist jedoch bei der verunglückten Person auf massiven Alkoholkonsum am Tage zuvor zurückzuführen. Hielte man Kreislaufschwäche konstant, würde man vielleicht eine Unabhängigkeit dieser Unfallart von Alkohol feststellen. So unmittelbar nachvollziehbar derartige kausale Relationen sind, so problematisch ist dieses einfach Modell für die Forderung nach der Betrachtung von bedingten Kontingenzen. Das Konstanthalten von B kann den Effekt von A auf C blockieren und zu dem Schluß führen, daß A in keinem ursächlichen Zusammenhang zu C steht (vgl. Waldmann, 1996). Die Regel der bedingten Kontingenz beim Kausallernen gilt also nur, wenn Kontextfaktoren nicht intermediär sind.

### *4.3* $A \rightarrow C \leftarrow B$ *-Modelle*

Eine solche Struktur impliziert beispielsweise $P(A, B) = P(A)P(B)$. Sobald aber der Effekt C festgelegt ist, sind A und B nicht mehr unabhängig: $P(A \mid B, C) \neq P(A \mid C)$ und $P(B \mid A, C) \neq P(B \mid C)$. Solche von Waldmann und Hagmayer (1998) als 'Gemeinsamer Effekt'-Modell bezeichnete Strukturen sind von kritischer Bedeutung. Häufig wird bei diesen Modellen der Art $A \rightarrow C \leftarrow B$ nur auf disjunktive multikausale Ursachenkonstellationen hingewiesen. Diese werden häufig mit dem Phänomen des sogenannten *explaining away* gleichgesetzt (vgl. Pearl, 1988). Unter *explaining away* ist dabei folgendes zu verstehen: In Fällen, in denen es zwei unterschiedliche alternative Ursachen gibt, führt die Annahme einer Ursache dazu, daß die andere Ursache weniger glaubhaft wird. Das folgende schlichte Beispiel macht deutlich, daß eine solche Struktur in einem BN durchaus auch adjunktive Ursachenkonstellationen abbilden kann. Nimmt man folgende probabilistische Beziehungen zwischen den Variablen A, B und C an: $P(C \mid A, B) = 0.95$; $P(C \mid \neg A, B) = 0.05$;
$P(C \mid A, \neg B) = 0.05$; $P(C \mid \neg A, \neg B) = 0.05$; alle übrigen bedingten Wahrscheinlichkeiten sind damit festgelegt; die Aprioriwahrscheinlichkeiten für die Variablen A und B werden auf 0.5 gesetzt. In diesem keineswegs abwegigen Beispiel, in dem erst das Zusammentreffen zweier Ursachen zu einer massiven Erhöhung der Wahrscheinlichkeit eines Effekts führt, in dem aber unter allen anderen Bedingungen der Effekt sehr unwahrscheinlich ist, tritt nun, vorausgesetzt C ist beobachtet worden, beim Feststellen der Ursache B keineswegs *explaining away* von A auf. Stattdessen steigt die aktualisierte Wahrscheinlichkeit von A in diesem Fall von 0.909 auf 0.95. Es sei an dieser Stelle ebenfalls darauf hingewiesen, daß der Begriff

[1] Auf die Zirkularität einer solchen Forderung weist Pearl (1996) hin.

*explaining away* auch in einer anderen begrifflich zu unterscheidenden Weise verwendet wird. So tritt neben der gerade dargestellten *spezifischen* Form eine *unspezifische* Form des *explaining away* auf. Unspezifisches *explaining away* kann sich bei der sozusagen „vorsichtigen“ Quantifizierung einzelner Ursachen ergeben, wenn eine Person annimmt, daß es noch andere mögliche Ursachen für einen Effekt gibt, ohne daß sie benennen kann, um welche es sich dabei handelt könnte (dies entspricht im Prinzip einer Verletzung der *closed world Annahme*).

Nichtsdestoweniger ist das Phänomen *explaining away* auch deshalb interessant, weil strukturelle Ähnlichkeiten zu den aus dem Bereich des Kausallernens bekannten Effekten der *cue competition* bzw. des *blocking* bestehen. Für eine Darstellung dieser Effekte sei auf Waldmann und Holyoak (1992) verwiesen. Die experimentellen Ergebnisse von Waldmann und Holyoak legen darüber hinaus nahe, daß Menschen bei der Bildung kausaler Hypothesen durchaus berücksichtigen, ob ein spezifischer Effekt durch einen oder mehrere konkurrierende Faktoren verursacht sein kann: „multiple alternative causes [thus] compete in diagnostic reasoning“ (Waldmann & Holyoak, 1992, p. 227). Einschränkend muß hervorgehoben werden, daß in der Untersuchung von Waldmann & Holyoak, 1992, deterministische Ursachen-Wirkungs-Zusammenhänge in der Lernphase präsentiert wurden. Ob *cue competition* in der von den Autoren vorhergesagten Weise auch unter nicht-deterministischen Beziehungen zwischen Ursache und Wirkung auftritt, bleibt eine offene Frage. Indirekte Evidenz für die Sensitivität gegenüber unterschiedlichen Kausalmodellen findet sich jedoch auch in einer Untersuchung zum *belief updating* von Plach (1998, 1999). Diese soll nun überblicksartig dargestellt werden.

Die Aufgabe für die Versuchspersonen bestand darin, die Wahrscheinlichkeit eines Verkehrsstaus in unterschiedlichen Verkehrssituationen zu schätzen und anschließend ihr Urteil auf der Grundlage neuer Information zu aktualisieren. Vor dem eigentlichen Experiment wurde ein BN gemäß den Vorannahmen der Versuchspersonen konstruiert. Die Variablen und die Struktur wurde aus Befragungen gewonnen. Das BN bestand aus 12 Variablen, von denen einige prädiktiv (z.B. Berufsverkehrszeit: ja/nein), andere diagnostisch (z.B. Verkehrsfunk meldet Stau: ja/nein), eingesetzt werden konnten, um die aktuelle Stauwahrscheinlichkeit zu schätzen. Im Experiment selbst wurde zunächst eine Quantifizierungsphase durchgeführt, deren Ziel es war, die für das festgelegte BN benötigten bedingten und unbedingten Wahrscheinlichkeiten zu erhalten. Aus der Struktur und den Quantifizierungen wurde deutlich, daß Versuchspersonen multikausale Interaktionen in flexibler Weise annehmen. Ein Beispiel für die im Sinne eines *explaining away* konzipierten Zusammenhangskomplexes ist in der Abhängigkeit der Variablen „Stau“ von den Variablen „Verkehrsunfall“ versus „Baustelle“ zu sehen. Ein Beispiel für eine adjunktive Ursachenkonstellation ist dagegen in der Abhängigkeit des „Verkehrsaufkommens“ von „Ballungszentrum“ und zeitlichen Variablen wie „Berufsverkehr“ und „Ferienbeginn“ zu sehen.

In der eigentlichen Experimentalphase wurde den Versuchspersonen unter systematisch variierten Bedingungen sukzessive Information präsentiert, die entweder direkt oder indirekt in Form eines diagnostischen oder prädiktiven Schlußes zu einer Veränderung der Stauwahrscheinlichkeit führen sollte. Die Ergebnisse des *updating* der Versuchspersonen wurden mit den Ergebnissen von Simulationen des BN verglichen. Die Simulationsbedingungen für das BN entsprachen dabei genau den einzelnen Informationsbedingungen, die für die Versuchspersonen galten. Die Ergebnisse lassen sich wie folgt zusammenfassen (für Details siehe Plach, 1998, 1999). (1) Die Updating-Ergebnisse

der Versuchspersonen stimmen in der Regel sehr gut mit den Vorhersagen des BN überein. Für das Hauptexperiment liegt die Korrelation zwischen Vorhersagen und Daten bei r=0.91. (2) Es lassen sich keine systematischen Verzerrungen von den Vorhersagen des BN Modells erkennen. (3) Neben dem BN wurden zwei Modelle getestet, die von der Kausalstruktur der Domäne abstrahieren. Es handelte sich dabei erstens um das Bayes Theorem unter Annahme der bedingten Unabhängigkeit aller Daten von der Zielvariable „Stau“ und zweitens um einen linearen Updating-Mechanismus, der von Carlson und Dulany (1988) vorgeschlagen wurde. Das BN sagt die Daten in beiden Fällen die Daten signifikant besser voraus als die Modelle, die die Kausalstruktur nicht berücksichtigen.

## 5. Schlußfolgerungen und Ausblick

Die Analyse der strukturellen Eigenschaften Bayesscher Netze und die generell gute Übereinstimmung dieser Eigenschaften mit empirischen Befunden haben gezeigt, daß diese Theorie ein interessantes Rahmenmodell für die Repräsentation und Verarbeitung von Kausalwissen an der Schnittstelle zwischen normativer und deskriptiver Theoriebildung bieten könnte. In zukünftigen Untersuchungen wird es notwendig sein zu überprüfen, inwieweit Personen in der Lage sind, ganz spezifische probabilistische Kausalmodelle zu akquirieren bzw. auf der Grundlage solcher spezifischer Modelle *Wissensaktualisierung* gemäß den Vorhersagen eines BN durchzuführen.

## 6. Literatur


Carlson, R.A., & Dulany, D.E. (1988). Diagnostic reasoning with circumstantial evidence. *Cognitive Psychology, 20,* 463-492.

Cartwright, N. (1989). *Nature's capacities and their measurement.* Oxford: Clarendon Press.

Eells, E. (1991). *Probabilistic causality.* Cambridge: Cambridge University Press.

Neapolitan, R.E. (1990). *Probabilistic reasoning in expert systems: theory and algorithms.* New York, NY.: Wiley.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Mateo, CA.: Morgan Kaufman.

Pearl, J. (1996). Structural and probabilistic causality. In D.R. Shanks, K.J. Holyoak, and D.L. Medin (Eds.), *The psychology of learning and motivation: Advances in research and theory,* (Vol. 34, pp. 393-436). New York, NY.: Academic Press.

Plach, M. (1998). *Prozesse der Urteilsrevision: Kognitive Modellierung der Verarbeitung unsicheren Wissens*. Wiesbaden: DUV.

Plach, M. (1999). Bayesian networks as models of human judgement under uncertainty: the role of causal assumptions in belief updating. *Kognitionswissenschaft, 8,* 30-39.

Russell, S. & Norvig, P. (1995). Artificial intelligence: a modern approach. Upper Saddle River, NJ.: Prentice-Hall.

Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning.* New York, NY.: Wiley.

Slovic, P. & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behavior and Human Performance, 6,* 649-744.

Spohn, W. (1993). Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein? In L.H. Eckensberger & U. Gähde (Hrsg.), *Ethische Norm und empirische Hypothese*, (S. 151-196). Frankfurt: Suhrkamp.

Waldmann, M. (1996). Knowledge-based causal induction. In D.R. Shanks, K.J. Holyoak, and D.L. Medin (Eds.), *The psychology of learning and motivation: Advances in research and theory*, (Vol. 34, pp. 47-88). New York, NY.: Academic Press.

Waldmann, M. & Holyoak, K.J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222-236.

Waldmann, M.R. & Hagmayer, Y. (1998). Die kognitive Konstruktion von Kausalität. *Dialektik, 2*, 101-114.

# Sind statistische Verfahren zur Kausalanalyse adäquate Modelle kausalen Denkens?

York Hagmayer und Michael R. Waldmann
Universität Göttingen

Wissenschaftler wie Laien stellen kausale Hypothesen auf und überprüfen diese mehr oder minder systematisch anhand von Erfahrungen. Trotz der zentralen Wichtigkeit dieses Prozesses sowohl für die Erweiterung des Wissens als auch für erfolgreiches praktisches Handeln ist noch wenig hierüber bekannt (Dunbar, 1993; Kuhn et al., 1995; Schauble, 1996).

Kausale Hypothesen können sich entweder auf einzelne Kausalzusammenhänge oder auf eine Kausalstruktur bestehend aus mehreren Kausalzusammenhängen beziehen. Bei einfachen Kausalhypothesen wird erstens gefragt, ob ein Kausalzusammenhang zwischen zwei Ereignissen überhaupt besteht, zweitens, ob dieser fördernd oder hemmend ist und drittens, wie stark dieser ist. Die Forschung zeigt, daß Personen die Kovariation zwischen der Ursache und dem vermeintlichen Effekt hierzu (Waldmann & Hagmayer, subm., Wasserman et al., 1993) oder daraus abgeleitete Größen wie Causal Power (Cheng, 1997) als Schätzer für den Kausalzusammenhang benutzen.

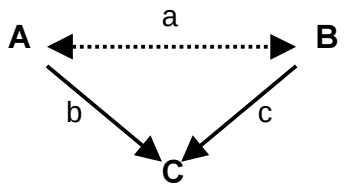
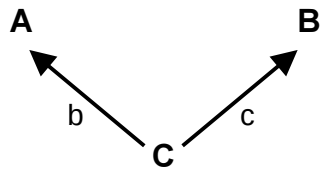
Bei Strukturhypothesen wird gefragt, ob drei oder mehr Ereignisse innerhalb eines kausalen Netzwerks durch eine spezifische Struktur verbunden sind. Bei drei Ereignissen ABC könnten zum Beispiel A und B C verursachen (Gemeinsamer-Effekt-Struktur) oder C könnte A und B hervorrufen (Gemeinsame-Ursache-Struktur) oder die drei Ereignisse bilden eine Kette.

Während bei zwei Ereignissen aufgrund ihrer Kovariation nicht auf die Richtung des Kausalzusammenhangs geschlossen werden kann (die zeitliche Abfolge ist ein notwendiges Zusatzkriterium), ist dies innerhalb einer Struktur möglich (ohne daß die zeitliche Abfolge bekannt sein muß). Wenn alle drei Ereignisse kovariieren, so ist dies mit sowohl mit einer Kette (A->B->C) als auch einer Gemeinsame-Ursache-Struktur (A<-B->C) verträglich. In beiden Fällen sollten A und C kovariieren, aber unkorreliert sein, wenn B konstant gehalten wird. Im Gegensatz dazu ist die Gemeinsamer-Effekt-Struktur (A->B<-C) distinkt. Hier würde man keine Kovariation von A und B erwarten, die durch den gemeinsamen Effekt B erklärbar ist. Wird allerdings B konstant gehalten, so sollte sich ein statistischer Zusammenhang von A und C zeigen. Diese Überlegungen machen deutlich, daß Strukturhypothesen keine Komplexe aus mehreren einfachen Kausalhypothesen sind und nicht auf diese reduziert werden können.

*Normative Modelle zur Prüfung kausaler Strukturhypothesen*

Es wurden in den vergangene Jahren und Jahrzehnten verschiedene Ansätze zur Prüfung kausaler Strukturhypothesen vorgeschlagen. Eine Gruppe von Verfahren sind Lineare Strukturgleichungsmodelle, z.B. LISREL (Jöreskog & Sörbom, 1993; Loehlin, 1998). Bei diesen werden aus der Strukturhypothese Gleichungen über den Zusammenhang zwischen den Ereignissen abgeleitet. Die Parameter dieser Gleichungen, d.h. die Stärke der Kausalzusammenhänge, werden aufgrund der beobachteten statistischen Zusammenhänge geschätzt.

Abb.1.: Strukturelle Implikationen von Lineraren Strukturgleichungsmodellen für verschiedene Kausalstrukturen

| Gemeinsamer-Effekt-Struktur | Gemeinsame-Ursache-Struktur |
|---|---|
|  |  |
| Abgeleitete Gleichungen:<br>$r_{AB} = a$<br>$r_{AC} = b + a*c$<br>$r_{BC} = c+a*b$ | Abgeleitete Gleichungen:<br>$r_{AB} = b*c$<br>$r_{CA} = b$<br>$r_{CB} = c$ |

r bezeichnet die zu erwartenden Korrelationen zwischen den Ereignissen,
a, b, c bezeichnen die Pfadkoeffizienten, b und c geben dabei die Stärke des jeweiligen Kausaleinflusses an.

Unter Annahme einer Struktur können über die beobachteten Korrelationen zwischen den Ereignissen die Pfadkoeffizienten berechnet werden. Wenn die vermutete Struktur zutrifft, müssen alle Gleichungen erfüllt sein. Genau dieser Sachverhalt ermöglicht eine Überprüfung der Strukturhypothese. Bezogen auf obiges Beispiel: Wenn eine Gemeinsame-Ursache-Struktur gegeben ist, dann muß die Korrelation zwischen A und B gleich dem Produkt aus den Korrelationen zwischen A,C und B,C sein. Ist das nicht der Fall, dann liegt keine Gemeinsame-Ursache-Struktur vor.

Eine zweite Gruppe von Modellen, welche die Überprüfung von Strukturhypothesen ermöglichen, sind Bayesianische Netzwerke (Jensen, 1996; Pearl, 1988). Die Hypothesenprüfung erfolgt durch die Prüfung von unbedingten und bedingten Unabhängigkeiten (conditional independence) zwischen den Ereignissen. So sind bei einer Gemeinsame-Ursache-Struktur die beiden Effekte abhängig voneinander (sie korrelieren). Wird allerdings C konstant gehalten, dann verschwindet diese Abhängigkeit, da die Kovariation zwischen A, B durch die Kovariation von A,C und B,C sowie Variation von C bedingt ist. Beim Gemeinsamer-Effekt-Modell sind die Vorhersagen genau entgegengesetzt. Hier führt das Konstanthalten des Effektes zu einer Abhängigkeit der Ursachen.

Abb. 2: Strukturelle Implikationen von Bayesianischen Modellen für verschiedene Kausalstrukturen

| Gemeinsamer Effekt-Struktur | Gemeinsame Ursache-Struktur |
|---|---|
| A, B sind unabhängig<br>$P(A\|B) = P(A)$<br><br>A, B sind abhängig bezogen auf C<br>$P(A\|C) \neq P(A\|C.B)$ | A, B sind abhängig<br>$P(A\|B) \neq P(A\|\sim B) \neq P(A)$<br><br>A, B sind unabhängig bezogen auf C<br>$P(A\|C) = P(A\|C.B)$ |

Bayesianische Modelle und Strukturgleichungsmodelle stehen nicht in einem grundsätzlichen Gegensatz zu einander, sie lassen sich vielmehr ineinander überführen (Pearl, 1998). Die Unabhängigkeitsannahmen, welche bei Bayesianischen Modellen direkt

geprüft werden, fließen bei Strukturgleichungsmodellen in die Bestimmung der Pfadkoeffzienten mit ein und werden hierdurch ebenfalls mitgeprüft. Ein Unterschied zwischen beiden Verfahren liegt darin, daß durch Prüfung der Unabhängigkeitsbeziehungen Strukturen induziert werden können. Die Gemeinsamer-Effekt-Struktur ist von allen anderen Strukturen eindeutig unterscheidbar und kann deshalb in lokalen Triaden allein durch Datenanalyse erschlossen werden. Lineare Strukturgleichungsmodelle werden in der Regel lediglich angewandt um vermutete globale Strukturen auf ihr Vorliegen hin überprüfen.

*Experimente*

In drei Experimenten wurde untersucht, ob das Vorgehen von Personen bei der Überprüfung von Strukturhypothesen dem von normativen Modellen äquivalent ist. Oder anders gesagt, die Fragestellung ist, ob Personen die Implikationen von Strukturhypothesen bei der Überprüfung dieser nutzen. Selbst wenn sie nur über ein grobes Verständnis der oben genannten Implikationen verfügen, sollten sie erwarten, daß zwischen den beiden Effekten einer Ursache eine Korrelation vorliegt und nur zwischen den beiden Ursachen eines Effektes kein Zusammenhang sein kann. Als Alternativhypothese haben wir angenommen, daß Personen die komplexen Strukturhypothesen in mehrere einfache Kausalhypothesen dekomponieren. Sie würden dann lediglich einzelne Kausalzusammenhänge auf ihr Vorhandensein überprüfen. Hierdurch ginge die Sensitivität für die Implikationen verschiedener Strukturen verloren.

In Experiment 1 wurden den Personen sukzessive vier verschiedene Strukturhypothesen über den Zusammenhang zwischen drei Ereignissen vorgegeben, welche sie anhand eines Datensatzes überprüfen sollten. Die präsentierten Daten sprachen entweder für eine Gemeinsamer-Effekt-Struktur, d.h. zwei der Ereignisse waren unabhängig voneinander, oder eine Gemeinsame-Ursache-Struktur, d.h. alle drei Ereignisse waren abhängig voneinander. Abbildung 3 zeigt die vier Hypothesen, die beiden verwendeten Datensätze sowie die Einschätzung der Versuchspersonen (Mittelwerte der Ratings auf einer Skala von 0 = ‚sicher nicht' bis +10 = ‚sicher ja'). Die Datensätze wurden bewußt extrem gewählt und die Korrelation zwischen den vermeintlichen Effekten von zu erwartenden 0.25 auf 0.50 überhöht, um diese salienter zu machen.

Abb. 3: Ergebnisse von Experiment 1 (N=48)

| Strukturhypothese<br>Daten | 1. A → C, B → C | 2. A ← C, B ← C | 3. A → B, A → C | 4. A ← B, A ← C |
|---|---|---|---|---|
| 1. Pro Gemeinsamer Effekt<br>A ..... .0 ..... B<br>.5 \ C / .5 | 6,6 | 7,6 | 2,0 | 2,7 |
| 2. Pro Gemeinsame Ursache<br>A ..... .5 ..... B<br>.5 \ C / .5 | 6,5 | 5,9 | 6,5 | 6,4 |

Die Ergebnisse deuten darauf hin, daß die Personen nicht sensitiv für die Implikationen verschiedener Strukturen sind (Hypothese 1 wird nicht von 2, und 3 nicht von 4 unterschieden). Vielmehr scheint es so, als ob sie das Vorhandensein der vermuteten Kausalzusammenhänge prüfen. Unter Hypothese 3 und 4 ist bei Datensatz 1 jeweils einer der vermuteten Zusammenhänge nicht vorhanden. Diese beiden Hypothesen werden als

signifikant weniger wahrscheinlich als Hypothese 1 und 2 beurteilt. Bei Datensatz 2 sind alle vermuteten Zusammenhänge vorhanden und die Hypothesen werden alle als gleich wahrscheinlich angesehen.

Das nicht normative Vorgehen wurde in einem zweiten Experiment ebenfalls gefunden. Die Personen mußten sich in diesem direkt zwischen einer Gemeinsame-Ursache-Struktur und einer Gemeinsamer-Effekt-Struktur entscheiden, wobei alle vermuteten Zusammenhänge jeweils vorhanden waren. Das Material entsprach den beiden Datensätzen und den Hypothesen 1 und 2 aus Experiment 1. Es zeigte sich keine Präferenz für die den Daten entsprechende Hypothese.

Für das dritte Experiment wurden mittels Bayesianischer Netze zwei Datensätze (Pro Gemeinsame-Ursache und pro Gemeinsamer-Effekt) entwickelt. Zusätzlich zu den statistischen Zusammenhängen waren in den jeweils hypothesenkonformen Daten die Ursachen notwendig für den Effekt. Trotz dieser starken Vereinfachung wurde das Ergebnis von Experiment 1 im wesentlichen repliziert. Die Differenzierung erfolgte zwischen Hypothesen, bei denen alle vermuteten Kausalzusammenhänge vorliegen, und solchen, bei denen dies nicht der Fall ist.

Von den insgesamt 148 Versuchspersonen in den drei Experimenten nahm bei der nachfolgenden mündlichen Befragung keine auf strukturelle Implikationen der vorgegebenen Hypothese bezug. Dies gibt einen weiteren Hinweis darauf, daß diese validen Kriterien nicht bewußt zur Entscheidung herangezogen werden.

*Diskussion und Ausblick*

In drei Experimenten konnten keine Belege dafür gefunden werden, daß Personen bei der Überprüfung von Strukturhypothesen auch nur in Ansätzen so vorgehen, wie dies normative Modelle zur Kausalanalyse tun. Sie nutzen nicht die Schlußfolgerungen, welche aus verschiedenen Strukturen über zu erwartende Abhängigkeiten zwischen den Ereignissen abgeleitet werden können. Die Frage, ob diese Modelle adäquate Modelle kausalen Denkens sind, muß daher wahrscheinlich verneint werden.

Dieser Befund überrascht vor allem deshalb, weil aus der Kategorisierungsforschung bereits mehrere Arbeiten vorliegen, welche eine Sensitivität für strukturelle Implikationen gefunden haben (Waldmann & Holyoak, 1992; Rehder, 1998). Dabei ist allerdings zu beachten, daß diese Arbeiten eine *implizite* Nutzung der Strukturen untersuchten. Versuchspersonen wurde eine Hypothese in bezug auf die Kausalstruktur der Merkmale innerhalb einer Zielkategorie vorgegeben. Der Einfluß der Strukturhypothese zeigte sich darin, daß das Lernen erschwert wurde, wenn die Lerndaten nicht der erwarteten Struktur entsprachen. Beim Hypothesenprüfen jedoch geht es um das *explizite* Wissen und Nutzen dieser Implikationen und hierfür ergab sich ein gegenteiliger Befund. Dissoziationen zwischen explizitem und impliziten Wissen sind aus der Lern- und Gedächtnisforschung hinlänglich bekannt. Ob dies auch beim Kausalen Denken der Fall ist, wird eine interessante Frage für die Zukunft sein.

Normative Modelle zur Prüfung von Kausalhypothesen gewinnen durch die scheinbare Unfähigkeit von Personen, zwischen verschiedenen möglichen Strukturen zu differenzieren, eine noch stärkere Bedeutung. Wenn wir nicht von uns aus diese Kriterien anwenden, so bedürfen wir um so mehr der Unterstützung durch diese Verfahren, wenn wir wissenschaftliche Erkenntnisse gewinnen wollen.

Literatur:

Cheng, P. (1997). From covariation to causation: A causal power theory.
*Psychological Review, 104*,367-405.

Dunbar, K. (1993). Concept discovery in a scientific domain.
*Cognitive Science*, *17*, 397-434.

Jensen, F.V. (1997). *An introduction to bayesian networks*. New York NY: Springer

Jöreskog & Sörbom (1993). Interactive LISREL 8.0. Chicago IL: Scientific Software

Kuhn, D. Garcia-Mila, M., Zohar, A., Andersen, Ch. (1995). Strategies of knowledge aquisition. *Monographs of the Society for Research in Child Development*, *60*.

Loehlin, J.C. (1998). Latent variable models: factor, path and structural analysis. Mahwah NJ: Lawrence Erlbaum.

Pearl, J. (1998). Graphs, causality and structural equation models. Technical Report. University of California Los Angeles.

Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo CA: Morgan Kaufmann.

Rehder, B. (1998). The essence of categories: The effect of underlying mechanism in induction, categorization and similarity. Unpublished PhD Thesis, University of Colorado.

Schauble, L. (1996). The development of scientific reasoning in knowledge rich contexts. *Developmental Psychology*, *32*, 102-119.

Waldmann, M. R. & Hagmayer, Y. (subsmitted). Pragmatic constraints on causal contingency assessments: The role of prior knowledge and processing effort.

Waldmann, M. R. & Holyoak, K. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psycology: General, 121*, 222-236

Wasserman, E.A., Elek, S.M., Chatlosh, D.L. & Baker, A.G. (1993). Rating Causal relations: Role of probability in judgements of response outcome contingency. *Journal of Experimental Psychology: Learning Memory and Cognition, 19*, 174-188.

# Kausale Effekte in der empirischen Sozialforschung

Christof Nachtigall, Rolf Steyer & Olivia Wüthrich-Martone

Institut für Psychologie der Universität Jena

**Zusammenfassung**

Ein wichtiger Schritt zur Präzisierung des Kausalitätsbegriffs ist das Konzept des *individuellen* und *durchschnittlichen kausalen Effekts.* Im Kontext der Psychologie spricht man von einem individuellen kausalen Effekt als dem Effekt, den ein Treatment auf eine Person hätte im Vergleich zur Nichtbehandlung. Allerdings besteht in der Regel nicht die Möglichkeit, individuelle kausale Effekte zu bestimmen, denn man kann eine Person nicht gleichzeitig behandeln und nichtbehandeln. Unter bestimmten Bedingungen, z. B. in randomisierten Experimenten ist es möglich, einen durchschnitt-lichen kausalen Effekt zu ermitteln. Dies ist der Durchschnitt aus den individuellen kausalen Effekten über alle Personen der betrachteten Population. Ausserhalb randomisierter Experimente besteht jedoch die Gefahr, dass konfundierende Variablen kausale Effekte verfälschen. Es wird ein Verfahren vorgestellt, mit dem Konfundierungen getestet werden können. Außerdem wird ein Weg aufgezeigt, wie im Falle einer vorliegenden Konfundierung dennoch auf kausale Effekte geschlossen werden kann.

## Einleitung

Es ist ein Hauptanliegen empirischer Forschung, kausale Schlussfolgerungen ziehen zu können. Im Bereich der Sozial- und Humanwissenschaften stellt sich ständig die Frage, ob eine bestimmte Maßnahme *wirklich* zu den angestrebten Konsequenzen führt. Kann beispielsweise in der Psychotherapie eine eingetretene Veränderung *kausal* auf vorausgehende psychologische Interventionen zurückgeführt werden? Oder umgekehrt: Kann aus dem Nicht-Eintreten einer erwarteten Wirkung auf die Nicht-Wirksamkeit der Intervention geschlossen werden?

In der Regel ist es ausserhalb randomisierter experimenteller Studien äußerst problematisch, solche Fragen zu beantworten. Es besteht grundsätzlich die Möglichkeit, dass beobachtete Effekte aufgrund von Störvariablen und nicht aufgrund der Behandlung eintreten. Betrachten wir ein Beispiel aus der Psychotherapieforschung. Wenn eine Gruppe von behandelten Patienten mit einer entsprechenden Gruppe von Personen der Warteliste verglichen wird, so kann der Behandlungseffekt (etwa die durchschnittliche Verbesserung der Befindlichkeit) bei der Therapiegruppe deutlich geringer ausfallen als in der Kontrollgruppe. Bedeutet dies, dass die "Warteliste" die effektivere Therapie ist? Sollte man alle Patienten besser untherapiert lassen? Möglicherweise erhalten aber Patienten mit der größeren Schwere der Symptomatik bevorzugt einen Therapieplatz, während leichtere Fälle auf der Warteliste verbleiben. In so einem Fall könnte die Schwere der Symptomatik eine konfundierende Variable sein, welche die Bewertung kausaler Effekte der Therapie verfälschen würde. Ziel dieses Beitrages ist es, Kriterien für kausale Effekte vorzustellen und mögliche empirische Überprüfungsmöglichkeiten zu demonstrieren.

## Kausale Effekte

Zurückgehend auf die Arbeiten von Neyman (1923/1990) spricht man von einem *individuellen kausalen Effekt* (engl. abgekürzt ICE) als dem Effekt, den eine Behandlung auf eine Person hätte im Vergleich zur Nichtbehandlung. Wenn z. B. der Effekt eines Treatments auf einer Skala von 0 bis 20 gemessen wird, so könnte eine Person $P$ nach der Behandlung den Wert $6$ und ohne Behandlung nur den Wert $2$ haben. Es gilt in diesem Fall $ICE_{Ther.\text{-}Kontr.}(P)=6-2=4$.

Das fundamentale Problem liegt aber darin, dass individuelle kausale Effekte in der Praxis kaum ermittelt werden können. Man kann eine Person nicht gleichzeitig behandeln und unbehandelt lassen. Unter Umständen ist es jedoch möglich, einen *durchschnittlichen kausalen Effekt* (engl. abgekürzt ACE) zu ermitteln. Dies ist der Durchschnitt aus den individuellen kausalen Effekten über alle

Personen der betrachteten Population. Er beschreibt die durchschnittliche Verbesserung unter der Therapiebedingung im Vergleich zur Wartelistenbedingung. Doch wie soll ein durchschnittlicher kausaler Effekt ermittelt werden, wenn die individuellen Effekte nicht bestimmt werden können?

Was ermittelt werden kann, sind Mittelwerte der Personen, welche tatsächlich therapiert werden, und der Personen, welche tatsächlich der Kontrollgruppenbedingung unterliegen. Im Rahmen von Wirksamkeitsstudien wird typischerweise eine Kriteriumsvariable Y für die Wirksamkeit eines Treatments bei einer Gruppe behandelter Personen erhoben. Dieser Mittelwert wird dann mit dem Mittelwert einer Kontrollgruppe verglichen. Betrachten wir Tabelle 1, wo die Werte der Kriteriumsvariable einer Population von acht Personen sowohl unter Therapie als auch unter Kontrollgruppenbedingung eingetragen sind.

| *Person* | *Therapie-bedingung* | *Kontrollgruppen-bedingung* | $ICE_{Ther.-Kontr.}$ | *Bedürftigkeit* |
|---|---|---|---|---|
| ***A*** | ***6*** | *2* | *4* | *hoch* |
| ***B*** | ***8*** | *5* | *3* | *hoch* |
| *C* | *6* | ***3*** | *3* | *hoch* |
| ***D*** | ***4*** | *2* | *2* | *hoch* |
| *E* | *12* | ***10*** | *2* | *niedrig* |
| *F* | *17* | ***13*** | *4* | *niedrig* |
| ***G*** | ***14*** | *11* | *3* | *niedrig* |
| *H* | *13* | ***10*** | *3* | *niedrig* |
| *Mittelwerte:* | $\mu_{Ther.ges.} = 10$ | $\mu_{Kontr.ges.} = 7$ | $ACE_{Ther.-Kontr.} = 3$ | |
| | $\mu_{Ther.} = \mathbf{8}$ | $\mu_{Kontr.} = \mathbf{9}$ | $PFE_{Ther.-Kontr.} = \mathbf{-1}$ | |

Tabelle 1: *Die Werte der Kriteriumsvariable bei acht Personen. Die fettgedruckten Werte zeigen an, welche Personen tatsächlich zur Therapie- und zur Kontrollgruppenbedingung zugewiesen wurden.*

Es wird ein Teil dieser Personen tatsächlich therapiert (die Personen *A, B, D* und *G*). Der Mittelwert von *Y* in der Therapiegruppe beträgt $\mu_{Ther.} = (6+8+4+14)/4 = 32/4 = 8$ (fettgedruckte Werte in der zweiten Spalte der Tabelle). Die anderen Personen *C, E, F* und *H* werden der Kontrollgruppe zugewiesen, hier ist der Mittelwert $\mu_{Kontr.} = (3+10+13+10)/4 = 36/4 = 9$ (fettgedruckte Werte der dritten Spalte der Tabelle). Die Differenz dieser beiden Mittelwerte ist *-1*. Eine solche Mittelwertdifferenz wird als *Prima Facie Effekt* (*PFE*) bezeichnet. Auf den ersten Blick erweist sich die Behandlung als weniger wirksam als die Kontrollgruppenbedingung. Der Prima Facie Effekt der Therapie- im Vergleich zur Kontrollbedingung ist mit $PFE_{Ther.-Kontr.} = -1$ negativ, behandelte Personen haben im Mittel einen niedrigeren Wert auf der Kriteriumsvariable als nicht-behandelte Personen. Doch eine kausale Interpretation dieses Ergebnisses würde zu vollkommen falschen Schlüssen führen. Betrachten wir die individuellen kausalen Effekte: Für *jede* Person ist der ICE, also die Differenz zwischen zweiter und dritter Spalte der Tabelle, positiv, bei Person *A* z. B. *+4*. Bei *jeder* Person würde die Behandlung eine positive Wirkung haben. Der durchschnittliche kausale Effekt ist der Durchschnitt der ICE und hat bezogen auf diese acht Personen den Wert $ACE_{Ther.-Kontr.} = +3$. Der ACE unterscheidet sich also nicht nur hinsichtlich der Größe von der Mittelwertdifferenz zwischen Therapie- und Kontrollgruppe, sondern der kausale Effekt weist sogar in die entgegengesetzte Richtung wie der Prima Facie Effekt. Solche zunächst höchst kontraintuitiven Phänomene sind unter dem Begriff des Simpson-Paradox bekannt geworden. Eine kausale Interpretation des Prima Facie Effects kann zu vollkommen falschen Schlüssen über die Wirksamkeit einer Behandlung führen. Daher ist es von großer Wichtigkeit, den Mechanismen einer solchen "kausalen Verfälschung" genauer kennenzulernen und Methoden zu entwickeln, mit denen eine eventuelle Verfälschung entdeckt werden kann.

**Wann liefern Vergleiche von Gruppenmittelwerten kausale Effekte?**

Trotz der im Beispiel beschrieben Risiken ist der Vergleich von Gruppenmittelwerten die am weitesten verbreitete Methode zur Interpretation empirischer Ergebnisse. Es stellt sich die Frage, unter welchen Bedingungen die Differenz von Gruppenmittelwerten[1] mit dem durchschnittlichen kausalen Effekt übereinstimmt. Mit Antworten auf diese Frage beschäftigen sich eine Reihe von Autoren (z. B. Holland & Rubin, 1983; Steyer u. a. 1996). Bekannt ist, dass im Falle von randomisierten Experimenten tatsächlich die Mittelwertdifferenz mit dem ACE übereinstimmt. Dies ist einer der wesentlichen Gründe für die überaus große Bedeutung experimenteller Forschung zur Ermittlung kausaler Zusammenhänge. Allerdings sind der Durchführung randomisierter Experimente in den Human- und Sozialwissenschaften oft enge Grenzen gesetzt. So läßt sich der sozio-ökonomische Status einer Person nicht experimentell manipulieren, um z. B. den kausalen Effekt auf die Lebenszufriedenheit zu ermitteln. Beschränkt man sich notgedrungen auf Beobachtungsstudien, indem man beispielsweise bereits bestehende Gruppen statushoher vs. statusniedriger Personen vergleicht, so liefert der Prima Facie Effekt keineswegs notwendigerweise kausale Effekte. Die Situation ist ähnlich in obigem Beispiel aus der klinischen Forschung: $PFE_{Ther.-Kontr.}$ und $ACE_{Ther.-Kontr.}$ stimmen keineswegs überein. Der Grund dafür wird aus der letzten Spalte von Tabelle 1 deutlich: Die Aufteilung der Personen zu Therapie- und Kontrollgruppe erfolgt nicht rein zufällig. Personen mit hoher Bedürftigkeit erhalten mit hoher Wahrscheinlichkeit auch einen Therapieplatz, Personen mit geringer Bedürftigkeit dagegen nicht, eine Situation, wie sie in der Praxis oft vorkommt oder zumindest aus ethischer Sicht oft vorkommen sollte. Es liegt keine *randomisierte Zuweisung* zu den verschiedenen Untersuchungsbedingungen vor. Die "Bedürftigkeit" der Personen erweist sich als eine *Konfundierung*. Damit ist gemeint, dass sowohl die Zuweisung zu den Untersuchungsbedingungen als auch die durchschnittlichen Ergebnisse unter den Untersuchungsbedingungen von einer weiteren Variable abhängen. In Tabelle 2 wird dies deutlich. Es werden drei von vier hochbedürftigen Personen der Therapiegruppe zugewiesen, die Zuweisungsquote zur Therapie insgesamt beträgt dagegen nur 50%. Außerdem hat die Kriteriumsvariable unter hochbedürftigen Personen in der Therapiebedingung den Mittelwert $\mu_{Ther.,hoch} = 6$, was niedriger ist als der Gesamtmittelwert $\mu_{Ther.} = 10$. Die Zuweisungswahrscheinlichkeiten und Mittelwerte sind in Tabelle 2 für die Gesamtpopulation und die beiden Teilpopulationen hoch- und niedrigbedürftiger Personen zusammengefasst.

| *Population* | *Zuweisungswahrscheinlichkeit* | | *Mittelwert der Kriteriumsvariable* | |
|---|---|---|---|---|
| | *Therapie* | *Kontrollgruppe* | *Therapie* | *Kontrollgruppe* |
| *gesamt* | 0.50 | 0.50 | 10 | 7 |
| *hochbedürftige* | 0.75 | 0.25 | 6 | 3 |
| *niedrigbedürftige* | 0.25 | 0.75 | 14 | 11 |

Tabelle 2: *Die Zuweisungswahrscheinlichkeit zu den Untersuchungsbedingung und die durchschnittlichen Ergebnisse unter den verschiedenen Bedingungen unterscheiden sich in den Teilpopulationen der hochbedürftigen und niedrigbedürftigen Personen von der Gesamtpopulation.*

In mathematisch präzisierter Form können nun Aussagen über Konfundierung und über den Zusammenhang mit kausalen Effekten gemacht werden. An dieser Stelle sollen jedoch nur anschauliche einige wichtige Aussagen dargestellt werden[2]. Allgemein gilt: Der Zusammenhang zwischen einem Kriterium Y und einer Treatmentvariable ist konfundiert, wenn es Teilpopulationen

[1] In diesem Beitrag beziehen wir uns immer auf wahre Mittelwerte der Population. Die wahren Gruppenmittelwerte und ihre Differenz können mit Stichprobenmittelwerten geschätzt werden.

[2] Für die formal korrekte Darstellung und die Beweise siehe Steyer et al. (1996). Für eine ausführliche Darstellung der Theorie individueller und durchschnittlicher kausaler Effekte siehe auch Steyer et al. (submitted). Beide Artikel können aus dem Internet unter *http://www.uni-jena.de/svw/metheval/Publik/index.html* heruntergeladen werden.

gibt, in denen sowohl die Zuweisungswahrscheinlichkeiten zu einer Treatmentbedingungen als auch die Mittelwerte unter dieser Treatmentbedingung von der Gesamtpopulation abweichen. Die im Therapiebeispiel dargestellte Situation ist also typisch für eine Konfundierung.

Kommen wir nun zur zentralen Frage hinsichtlich des Vergleichs von Gruppenmittelwerten zur Ermittlung des durchschnittlichen kausalen Effekts. Konkret lautet die Frage: Unter welchen Bedingungen stimmen PFE und ACE überein, und wann unterscheiden sie sich. Die entscheidende Aussage hierzu lautet:

*Theorem 1: Unterscheidet sich eine Differenz von Gruppenmittelwerten (PFE) und der durchschnittliche kausale Effekt (ACE), so muss es eine Konfundierung geben.*

*Umgekehrt gilt: Falls es keine Konfundierung gibt, so stimmt die Differenz von Gruppenmittelwerten (PFE) mit dem durchschnittlichen kausalen Effekt (ACE) überein.*

Es sei daran erinnert, dass in diesem Beitrag nur von "wahren" Gruppenmittelwerten, also von Populationsmittelwerten, die Rede ist. Diese sind aufgrund von Daten zu schätzen. Der PFE wird beispielsweise durch Stichprobenmittelwerte von Treatment- und Kontrollgruppe geschätzt. Die Abwesenheit einer Konfundierung stellt sicher, dass diese Schätzung aufgrund von Stichprobenmittelwerten kausal interpretiert werden kann, in dem Sinne, dass mit dieser Schätzung auch der durchschnittliche kausale Effekt geschätzt wird. Für die empirische Sozialforschung ist es daher wichtig, Verfahren an die Hand zu bekommen, mit denen Konfundierungen erkannt werden können.

**Konfundierungstests**

Sind eine Kriteriumsvariable $Y$ und eine Treatmentvariable $X$ gegeben (im Beispiel hat $X$ zwei Werte, nämlich *Therapie*- und *Kontrollgruppe*), so ist nach potentiell konfundierenden Variablen $W$ Ausschau zu halten. Eine solcher Kandidat für Konfundierung teilt die Population in Subpopulationen $W = w$. Im Beispiel ist $W$ die Bedürftigkeit der Personen und die Subpopulationen sind durch $W = hochbedürftige$ und $W = niedrigbedürftige$ Personen gegeben. Das folgende Theorem nennt empirisch prüfbare Bedingungen für eine Konfundierung.

*Theorem 2: Gibt es eine Treatmentbedingung $X=x$ und eine Subpopulation $W=w$, so daß für dieses $x$ und dieses $w$ die Bedingungen $\mu_{X=x} \neq \mu_{X=x,\ W=w}$ und $P(X=x \mid W=w) \neq P(X=x)$ erfüllt sind, dann liegt eine Konfundierung vor.*

Dabei ist $\mu_{X=x}$ der Mittelwert der Kriteriumsvariable in der Treatmentbedingung $X=x$ und $\mu_{X=x,\ W=w}$ ist der Mittelwert der Kriteriumsvariable unter $X=x$ in der Subpopulation $W=w$. Alle beteiligten Größen sind aus empirischen Daten schätzbar. Auf diese Weise kann eine mögliche Konfundierung mit Signifikanztests ermittelt werden. Wie man in Tabelle 2 ablesen kann, liegt dort eine Konfundierung vor. Bei hoch- wie niedrigbedürftigen Personen unterscheiden sich sowohl die Zuweisungswahrscheinlichkeiten als auch die Mittelwerte von denen der Gesamtgruppe. Weiterhin wird durch Theorem 1 und 2 deutlich, warum in randomisierten Experimenten PFE und ACE übereinstimmen müssen. Randomisierung bedeutet, dass die Zuweisung zu einer Treatmentbedingung rein zufällig erfolgt, es gilt also $P(X=x) = P(X=x \mid W=w)$ für alle Treatmentbedingungen $x$ und alle Subpopulationen $w$, daher kann keine Konfundierung vorliegen.

**Adjustierung von Mittelwerten**

Führt ein Konfundierungstest zu dem Ergebnis, dass eine Konfundierung vorliegt, so ist nicht mehr sichergestellt, dass die Differenz von Gruppenmittelwerten kausal interpretiert werden kann. Es ist jedoch unter Umständen möglich, trotz einer vorliegenden Konfundierung kausale Effekte zu berechnen. Dies geschieht durch eine geeignete Adjustierung der Gruppenmittelwerte. Betrachten wir wieder das Beispiel. In Tabelle 1 finden sich in der letzten Zeile die Mittelwerte der Kriteriums-

variablen bezogen auf die Gesamtpopulation (in der Therapiebedingung $\mu_{Ther.ges} = 10$, in der Kontrollbedingung $\mu_{Kontr.ges.} = 7$). Ihre Differenz ist der gesuchte durchschnittliche kausale Effekt. Empirisch zugänglich sind jedoch zunächst nur die Mittelwerte der tatsächlich therapierten Personen und der tatsächlich der Kontrollgruppe zugewiesenen Personen (die Werte sind $\mu_{Ther.} = 8$ *und* $\mu_{Kontr.} = 9$). Das Vorgehen besteht nun dahin, die Konfundierung durch die Bedürftigkeit *W*, die zu verfälschten Gruppenmittelwerten führte, gewissermaßen rückgängig zu machen. Konkret werden aus den Mittelwerten in den Subgruppen durch geeignete Gewichtung die gesuchten Mittelwerte $\mu_{Ther.,ges.}$ und $\mu_{Kontr.ges.}$ berechnet. Konkret muß der Ausdruck

$$\mu_{X=Ther.adj.} = \mu_{X=Ther.\ W=hoch} \cdot P(W=hoch) + \mu_{X=Ther.\ W=niedrig} \cdot P(W=niedrig)$$

berechnet werden. Im vorliegenden Fall beträgt die Quote der hoch- und niedrigbedürftigen jeweils 50%, es ist also $P(W=hoch)=P(W=niedrig)=0.5$. Aus Tabelle 2 können die Subgruppenmittelwerte $\mu_{X=Ther.,W=hoch} = 6$ und $\mu_{X=Ther.,W=niedrig} = 14$ entnommen werden. Insgesamt ergibt sich ein adjustierter Mittelwert von $\mu_{X=Ther.adj} = 10$. Dieser stimmt mit dem Gesamtmittel der Population $\mu_{Ther.ges.}$ überein. Entsprechend wird $\mu_{X=Kontr.adj} = 7 = \mu_{X=Kontr.ges}$ berechnet. Aufgrund der adjustierten Mittelwerte kann daher der ACE durch $\mu_{X=Ther.adj} - \mu_{X=Kontr.adj} = 3$ ermittelt werden.

Im allgemeinen Fall einer Kriteriumsvariable Y, einer Treatmentvariable X und einer konfundierenden Variable *W* mit Werten $w_1,\ldots w_n$ erfolgt die Adjustierung durch den Ausdruck

$$\mu_{X=x,adj.} = \sum_{i=1}^{n} \mu_{X=x,W=w} P(W=w)$$

Alle verwendeten Größen sind empirisch schätzbar. Es ist möglich, bei Vorliegen einer konfundierenden Variable W den durchschnittlichen kausalen Effekt zu bestimmen. Bedingung ist, dass es in den durch die Werte von W gebildeten Subgruppen keine weiteren Konfundierungen mehr gibt. In diesem Fall gilt

$$ACE_{x1\text{-}x2} = \mu_{X=x1,\ adj.} - \mu_{X=x2,\ adj.}.$$

Im Falle weiterer Konfundierungen besteht die Möglichkeit, diese durch Testung zu ermitteln und durch weitere Adjustierungen den durchschnittlichen kausalen Effekt zu berechnen.

## Diskussion

In der empirischen Sozialforschung ist es häufige Praxis, Differenzen von Gruppenmittelwerten kausal zu interpretieren. Beispiele zeigen jedoch, dass solche Mittelwertdifferenzen ausserhalb randomisierter Experimente keineswegs immer mit dem durchschnittlichen kausalen Effekt übereinstimmen, ja es müssen die Effekte noch nicht einmal das gleiche Vorzeichen haben. In diesem Beitrag wurden Bedingungen aufgezeigt, wann eine Mittelwertdifferenz zwischen Therapie- und Kontrollgruppe einem durchschnittlichen kausalen Effekt entspricht und daher als "kausal unverfälscht" betrachtet werden kann. Zur Ermittlung möglicher verfälschender Einflüsse wurde ein Verfahren vorgestellt, mit dem man Variablen daraufhin testen kann, ob es sich dabei wirklich um konfundierende Variablen handelt. Selbst im Falle einer vorliegenden Konfundierung ist es unter Umständen möglich, durchschnittliche kausale Effekte durch Adjustierung zu berechnen. Allerdings unterliegen die vorgeschlagenen Verfahren unterliegen praktischen Einschränkungen. Zunächst müssen potentiell konfundierenden Variablen auch erfasst werden. Konfundierungstest benötigen Parameterschätzung in Subpopulationen, was bei kleinen Stichproben problematisch werden kann. Ist eine konfundierende Variable *W* ermittelt, so wird zur Adjustierung die Verteilung von *W* benötigt. Das Ergebnis der Adjustierung liefert aber nur dann einen kausalen Effekt, wenn keine weitere Konfundierung in den Subpopulationen *W=w* vorliegt. Dies kann zwar wiederum mit dem Konfundierungstest geprüft werden, stößt aus praktischer Sicht erst recht auf die erwähnten Probleme mit kleinen Stichproben.

Die Idee der Adjustierung von Effekten zur Bereinigung von konfundierenden Einflüssen ist selbstverständlich nicht neu. Techniken der Parallelisierung von Stichproben, der Kontrolle von Störvariablen und deren "Bereinigung" werden in Versuchsplanung und statistischer Datenanlyse

häufig verwendet (vgl. z. B. Cochran & Rubin, 1973, Holland, 1986). Unklar bisher ist jedoch, welche der bestehenden Adjustierungsmethoden aus kausaltheoretischer Sicht sinnvoll sind und welche nicht. So konnten Wüthrich-Martone et al. (1999) zeigen, dass *keine* der im Rahmen der vom Statistikprogrammpaket SPSS angebotenen Adjustierungen bei der nichtorthogonalen Varianzanalyse zu einer Testung kausaler Effekte führt. Der hier beschriebene kausaltheoretische Rahmen ist daher auch eine nützliche Basis, um gezielt Versuchsplanungs- und Datenanalysetechniken zur Ermittlung kausaler Effekte auswählen und weiterentwickeln zu können.

**Literatur**


Cochran, W. G. & Rubin, D. B. (1973): Controlling Bias in Observational Studies: A Review. *Sankhya* A, 35, 417-446.

Holland, P.W. & Rubin, D.B. (1983). On Lord's Paradoxon. In H. Wainer & S. Messick (eds*.), Principles of modern psychological measurement* (pp 3-25). Hillsdale, NJ: Erlbaum.

Holland, P.W. (1986). Statistics and Causal Inference (with Comments). *Journal of the American Statistical Association,* 81, 945-970.

Neyman, J. (1923/1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. *Statistical Science*, 4, 465-480.

Steyer, R., Gabler, S., and Rucai, A..A. (1996). Individual Causal Effects, Average Causal Effects, and Unconfoundedness in Regression Models. In F. Faulbaum & W. Bandilla (eds.), *SoftStat '95. Advances in statistical software 5* (pp 203-2010). Stuttgart: Lucius & Lucius.

Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C. & Buhl, T. (submitted). Causal Regression Models I: Individual and Average Causal Effects. *Methods of Psychological Research-Online.*

Wüthrich-Martone, O., Steyer, R., Nachtigall, C. & Suhl, U. (1999). Causality, Confounding and Unbalanced Analysis of Variance. In: H. Friedl, A. Berghold & G. Kauermann (eds.)*, Statistical Modelling,. Proceedings of the 14th International Workshop on Statistical Modelling.* Graz, Austria, July 1999.



Korrespondenzadresse:

Dr. Christof Nachtigall
Institut für Psychologie der Universität Jena
Steiger 3 Haus 1
D-07743 Jena

e-mail: Christof.Nachtigall@uni-jena.de
Tel.: 03641-945234
Fax: 03641-945232
http://www.uni-jena.de/svw/metheval/

# Die Mathematisierung der Wissenschaften und die Beobachtbarkeit von Kausalrelationen

*Uwe Oestermeier (uwe_oestermeier@diff.uni-tuebingen.de)*

Seit langem bestehen in Philosophie, Wissenschaftstheorie und Psychologie unterschiedliche Auffassungen darüber, ob Kausalrelationen im Einzelfall beobachtbar sind oder nicht. Dieser Dissens läßt sich auflösen, wenn man davon ausgeht, dass in der Debatte unterschiedliche alltagssprachliche und normative Beobachtungsbegriffe Verwendung finden. Es wird gezeigt, dass ein normativer Beobachtungsbegriff, der strikt zwischen Beobachtung und kausaler Interpretation trennt sowie die Intersubjektivität, Objektivität und Präzision von Beobachtungen fordert, kognitionspsychologisch gesehen Anforderungen an das Erkenntnissubjekt stellt, die in der Praxis erst mit der Mathematisierung der Wissenschaften erfüllt werden konnten. Diese Mathematisierung wird als eine Erweiterung kognitiver Fähigkeiten durch externe Repräsentationen und kognitive Artefakte beschrieben.

In Philosophie, Wissenschaftstheorie und Psychologie besteht seit langem ein Dissens darüber, ob Kausalrelationen beobachtbar sind oder nicht. Die Position, dass Ursache-Wirkungs-Beziehungen als solche im Einzelfall prinzipiell nicht beobachtet sondern nur inferiert werden können, wird in der Regel auf David Hume zurückgeführt.[1] Hume zufolge kann z.B. beim Zusammenstoß zweier Billiardkugeln lediglich die *zeitliche Priorität* der Ursache vor der Wirkung sowie die *raumzeitliche Nähe* von Ursache und Wirkung beobachtet werden. Die Notwendigkeit, mit der die Wirkung (Bewegung) auf die Ursache (Stoß) folgt, ist hingegen eine Zutat, die aufgrund vorhergehender Erfahrungen in einzelne Ereignissequenzen hinein interpretiert wird. Ein Großteil der modernen Wissenschaftstheoretiker hält dies für eine wichtige und richtige Entdeckung Humes (Stegmüller 1983) und viele Psychologen haben sich der Humeschen Auffassung angeschlossen:

"Because causal relations are neither observable nor deducible, they must be induced from observable events" (Cheng 1997, S. 367).

"It is logically impossible to observe causal processes per se" (Ross & Fletcher 1985, S. 112).

In direktem Widerspruch dazu steht die gestaltpsychologische Traditionslinie der "phänomenalen Kausalität", die davon ausgeht, dass gerade elementare mechanische Vorgänge, wie Hume sie in seinem Billiardbeispiel betrachtet, direkt beobachtet werden können, wenn sie näher zu spezifizierenden Gestaltgesetzen gehorchen (Michotte 1982).[2]

Eine naheliegende Erklärung für den Dissens besteht darin, dass hier unterschiedliche Beobachtungsbegriffe Anwendung finden. Alltagssprachlich ist es jedenfalls völlig korrekt, Beobachtungs- und Wahrnehmungsverben mit Beschreibungen zu verknüpfen, die eine Kausalbeziehung implizieren:[3]

---

[1] Hume war sicherlich der einflussreichste Autor in dieser Hinsicht, auch wenn fast alle Einzelaspekte seiner Kausalanalyse bereits vorher von islamischen Philosophen, Scholastiker und französischen Okkasionalisten vorweggenommen worden waren (vgl. Weinberg 1973).

[2] Diese psychologische Debatte hat ein sprachphilosophisches Pendant. Dabei geht es um die Frage, ob jede singuläre Kausalaussage ein allgemeines Gesetz impliziert oder allgemeine Kausalaussagen Generalisierungen singulärer Kausalaussagen sind, die ihrerseits keine Verallgemeinerung implizieren (vgl. Ducassse 1993, Mackie 1974).

[3] Wer seiner eigenen Sprachintuition mißtraut, kann in Zeitungsarchiven zahlreiche Beispiele dafür finden, dass der allgemeine Sprachgebrauch tatsächliche solche Konstruktionen zuläßt.

"Ich habe gesehen, wie die Katze die Milch aufgeschleckt hat."

"Er hat doch beobachtet, dass der kleine Junge das Rad umgestoßen hat."

Dem Alltagsverständnis zufolge gibt es also spezifische Kausalrelationen, die der Beobachtung zugänglich sind. Aus Sicht eines Humeaners zeigt dies jedoch nur, dass der alltägliche Sprachgebrauch zu unscharf ist, um die Beobachtbarkeit von Kausalrelationen präzise analysieren zu können.

Welche Position letztlich die besseren Argumente auf ihrer Seite hat, soll hier nicht weiter untersucht werden. Ich gehe hier lediglich von der Annahme aus, dass im Gefolge Humes viele moderne Wissenschaftstheoretiker und Psychologen einen *normativen* Beobachtungsbegriff verwenden. Meine Hauptthese ist, dass dieser Beobachtungsbegriff an kultur- und wissenschaftshistorische Voraussetzungen der Darstell- und Messbarkeit von Beobachtungen geknüpft ist, die es überhaupt erst ermöglichen, systematisch zwischen einer reinen Beobachtung und einer kausalen Interpretation der Beobachtung zu unterscheiden. Dieser Beobachtungsbegriff hat sich mit der Mathematisierung der Naturwissenschaften entwickelt und ist insofern normativ, als er fordert, dass Beobachtungen

- von den daraus zu ziehenden kausalen Schlußfolgerungen strikt zu trennen sind;
- möglichst unabhängig von dem beobachtenden Subjekt sein sollen, d.h. im Idealfall *intersubjektiv*, *objektiv* und *präzise* sein sollen.

Diese Forderungen können Menschen, die nur über ihre Alltagssprache und ihre Sinnesorgane verfügen, nicht erfüllen. Der normative Beobachtungsbegriff setzt also stillschweigend eine kulturelles Milieu mit externen Repräsentationen (wie z.B. Kalküle zur Beweisführung und Tabellen zur Codierung von Beobachtungsdaten) und kognitive Artefakten (wie z.B. Messgeräte) voraus, mit denen der Mensch seine Schlußfolgerungs- und Wahrnehmungsfähigkeiten erweitert. Diese Darstellungsformen und Artefakte gehören zu unserem für selbstverständlich gehaltenen kulturellen Hintergrund, beruhen aber auf Entdeckungen und Erfindungen, die jeweils für sich genommen keineswegs zwangsläufig und naheliegend waren.

## Die Entwicklung einer neutralen Beobachtungssprache

Beispiele dafür, dass in der Alltagssprache Beobachtungen mittels kausaler Ausdrücke beschrieben werden, lassen sich beliebig aneinanderreihen.[4] In der heutigen Statistikausbildung wird hingegen strikt zwischen deskriptiver Statistik und Inferenzstatistik unterschieden. Dies war nicht immer so. In psychologischen Zeitschriften vor ca. 1935 wird dieser Unterschied fast gar nicht beachtet. Dies ändert sich erst mit der Aufnahme des t-Tests und der Varianzanalyse in das Arsenal der Standardmethoden der Psychologie (Gigerenzer 1987). Man mag dies für einen der vielen Zufälle in der Geschichte der Psychologie halten, aber dahinter steckt m.E. mehr: ein systematischer Übergang von normalsprachlichen Beschreibungen und Interpretationen zu quantitativen Beschreibungen mit formalisiertem Inferenzverfahren. Dieser Übergang war und ist auch in anderen Wissenschaften zu beobachten, die sich an dem Ideal der neuzeitlichen Physik orientieren.

Die kognitiven Voraussetzungen einer konsequenten Trennung von Beschreibung und Interpretation werden erst deutlich, wenn man alle wissenschaftlichen Hilfsmittel ausblendet: Publikationen und Laboratorien, Tabellen und Messgeräte, Graphen und mathematische Berechnungsverfahren etc. Orale Kulturen, die völlig ohne diese Hilfsmittel auskommen, sterben aus, aber gerade sie eröffnen interessante Vergleichsmöglichkeiten. Eine charakteristische Eigenschaft oraler Kulturen scheint zu sein, dass kausale Erklärungen primär narrativ

[4] Eine strikte Trennung von Wahrnehmung und Schlußfolgerung ist auch innerhalb der Psychologie nicht selbstverständlich, wie die zahlreichen Ansätze belegen, die Wahrnehmung als Inferenzprozesse deuten (z.B. Rock 1983).

tradiert werden, während in literalen Kulturen neben narrativen Genres auch argumentativ-logische Text- und Diskursformen entwickelt wurden (Donald 1991, Lindberg 1992). Schrift und Buchdruck sind damit die Voraussetzung für die wissenschaftliche Revolution, die auch als Übergang vom mythischen zum wissenschaftlichen, vom konkreten zum abstrakten, vom abergläubischen zum rationalen Denken beschrieben wurde. Solche Schlagworte und Dichotomien sind natürlich immer mit Vorbehalt zu betrachten, aber niemand kann bezweifeln, dass es unterschiedliche Niveaus des Wissens und mithin des kausalen Wissens gibt. Verschiedene Textsorten unterstützen dabei die strikte Trennung von Beobachtung und Interpretation in sehr unterschiedlicher Weise.

**Narrative Texte.** Wie in gesprochenen Erzählungen suggeriert auch in geschriebenen Texten die lineare Struktur der zeitlich nacheinander zu lesenden Wörter bereits eine kausale Ordnung. Die Sätze "Hans wurde krank und ging zum Doktor" und "Hans ging zum Doktor und wurde krank" sind in dieser Hinsicht völlig unterschiedlich zu interpretieren, ohne dass ein expliziter Hinweis auf zeitliche und kausale Ordnung vorliegt. Zuhörer und Leser ziehen die entsprechenden kausalen Schlußfolgerungen stillschweigend und automatisch (Warren, Nicholas, & Trabasso, 1979). In der Alltagssprache werden Ereignisfolgen also nicht erst in einer neutralen Beobachtungssprache beschrieben um dann in weiteren definierten Schlußfolgerungsschritten Kausalurteile daraus abzuleiten, sondern die alltagssprachliche Beschreibung von Vorgängen impliziert bereits immer eine kausale Interpretation dieser Vorgänge. Strawson (1985) weist in diesem Zusammenhang darauf hin, dass unser Alltagsvokabular für Kausalverhältnisse ähnlich differenziert ist, wie unser Vokabular für materielle Gegenstände. So wie wir eine Fülle von Objektkategorien haben (Möbel, Berge, Geschirr, Fahrzeuge, Nahrungsmittel etc.) haben wir eine Fülle von transitiven Verben (stoßen, ziehen, heben, biegen, kippen, drücken etc.) und Dispositionsprädikaten (wasserlöslich, mutig etc.), die beschreiben, wie etwas oder jemand auf seine Umgebung einwirkt. Diese Ausdrücke gehören zu unserem Beobachtungsvokabular, d.h. wir beschreiben Ereignisfolgen spontan und unmittelbar nach Einzelbeobachtungen als kausal, ohne die Ausdrücke "Ursache" und "Wirkung" zu benutzen.

**Normalsprachliche argumentative Texte.** Das kausale Alltagswissen bildet den Hintergrund oder Rahmen für den wissenschaftlichen Sprachgebrauch mit seinen expliziten Definitionen und Kriterien. Eine explizite argumentativ-logische Wissenschaftsprache konnte sich jedoch erst mit der Erfindung der Schrift ausbilden. Das heißt nicht, das in der Wissenschaftsprosa die basalen narrativen Prinzipien völlig außer Kraft gesetzt sind, sondern lediglich, dass in der Wissenschaftssprache in einem weit höheren Maß logische Strukturen durch Schlüsselwörter wie "weil", "somit", "deshalb", "These", "Prämisse", "Argument" explizit gekennzeichnet werden. Der Unterschied zu narrativen Genres ist jedoch graduell. Zwar kann man prinzipiell mittels des üblichen Wissenschaftsvokabulars alle nötigen epistemischen Unterscheidungen auch explizit benennen, aber Beschreibungen und Beobachtungen werden in der Praxis trotzdem oftmals nicht von theoretischen Überlegungen und Argumenten abgegrenzt (Gigerenzer 1987).

**Idealsprachliche argumentative Texte.** Ein Bewußtsein für logische Strenge und damit für den Unterschied von Prämissen und Konklusionen ist erst mit der Entwicklung rigider Beweisformen in Geometrie, Algebra und Logik entstanden. In all diesen Beweisformen spielt die lineare Ordnung der Erzählzeit der normalen Sprache keine Rolle mehr.

Vor diesem Hintergrund kann man geometrische, algebraische und logische Beweisformen als intellektuelles Zwangskorsett ansehen, die ein Wissenschaftler freiwillig benutzt, um sich zur Explizierung seiner Voraussetzungen und der Kontrolle der Beweisschritte zu zwingen. In der historischen Entwicklung kann man dabei eine Tendenz zu immer abstrakteren und kompakteren Repräsentationen erkennen. In einem modernen Kalkül repräsentiert je-

des Zeichen einen Begriff und nicht einen Vokal oder eine Silbe, wie in den normalen Schriftsprachen.[5] Nur in literalen Kulturen ist dieser Abstraktionsgrad des argumentativen Denkens erreicht worden (Olson 1996). Die Externalisierung von Gedanken auf Papier und ähnlichem scheint eine notwendige Bedingung dafür zu sein, es gibt jedenfalls keine oralen Gegenstücke zu formalen logischen, algebraischen, geometrischen und statistischen Schlußfolgerungsformen (Goody 1987).

Der kognitive Effekt von Notationen, wie sie in Beweisen benutzt werden, ist nicht zu unterschätzen. Wenn man statt normalsprachlicher verbaler Ausdrücke, die fast immer kausale Konnotionen haben, bedeutungslose Symbole verwendet, erliegt man auch nicht der Gefahr, unwillkürliche kausale Schlüsse und Verallgemeinerungen zu ziehen. Im Gegensatz zur Alltagssprache, in der die Asymmetrie von Ursache und Wirkung, Agens und Patiens tief verwurzelt ist (Fiedler, Semin & Finkenauer, 1994), erfordern formale Darstellungen zusätzliche Interpretationsschritte und Erläuterungen, um die Asymmetrie von Ursache und Wirkung zum Ausdruck bringen zu können. Die Gleichung $F = m \cdot a$ abstrahiert z.B. völlig von der Tatsache, dass im Alltag die Kraft als Ursache der Beschleunigung angesehen wird und nicht umgekehrt.

Ein Charakteristikum mathematischer Darstellungsformen ist dabei ihre Doppelfunktion als Repräsentations- und Operationsmedium. Die Symbole sind nicht nur Stellvertreter für Aussagen und Annahmen, sie sind gleichzeitig ein Medium mit dem man gemäß der erlaubten Beweisregeln definierte Operationen durchführen kann. Dadurch zwingt die Notation dazu, verborgene Prämissen zu explizieren und jeden Beweisschritt in nachvollziehbarer Weise durchzuführen.[6]

Logische und statistische Kalküle werden heutzutage nur von wenigen Experten beherrscht, lediglich das schriftliche Rechnen wurde in Europa mit der Erfindung des Buchdrucks zu einer allgemein verbreiteten kognitiven Fähigkeit (Menninger 1969). Mit dem algebraischen Rechnen und der Erfindung der analytischen Geometrie wurden quantitative Argumente zum Ideal der Naturwissenschaften. Quantitative und formale Argumentationsmuster übertreffen die normale Sprache in Präzision, Uniformität und Allgemeinheitsgrad und sind undenkbar in ausschließlich oralen Kulturen. Es ist deshalb wohl kein Zufall, dass der Begriff des quantitativen Naturgesetzes, der im Gegensatz zur Alltagssprache von Inhalten und Kontexten abstrahiert, nur in Gesellschaften mit elaboriertem mathematischen Wissen auftritt (Kelsen 1941).

### Wissenschaftliche Qualitätsstandards für Beobachtungen

In Alltagssituationen sind kausale Probleme (Wer ist Schuld am Unfall? Wer hat das Glas umgeschmissen? Woher kommt der Ausschlag?) meistens nicht durch wiederholte experimentell kontrollierte Beobachtungen zu entscheiden. Man kann Unfall- und Krankheitsverläufe nicht systematisch variieren. Des weiteren gibt es in vielen Fällen keine oder nur wenige

---

[5] Aristoteles war der erste, der in seiner Syllogistik Buchstaben als Variablen für Terme benutzte und damit komplett vom Inhalt der Argumente abstrahierte. In der Algebra wurde diese symbolische Variablennotation von Vieta eingeführt und erst mit Descartes kam es zu der heute geläufigen Praxis die letzten Buchstaben des Alphabets x,y und z für Unbekannte und die anderen für bekannte Werte zu nutzen.

[6] Rumelhardt (1989) hat darauf hingewiesen, dass unser Gehirn primär eine Mustererkennungsapparat ist und relativ schlecht für abstrakte Schlußfolgerungen ausgerüstet ist. Mit externen Notationen können abstrakte Schlußfolgerungsprobleme in Mustererkennungsprobleme übersetzt werden. Die Pioniere der modernen Logik waren sich der weitreichenden kognitiven Auswirkungen bewußt: "... by the aid of symbolism, we can make transitions in reasoning almost mechanically by the eye, which otherwise would call into play the higher faculties of the brain." (Whitehead, 1911, zitiert nach Copi, 1968, p. 213) Copi (1968) hat Whitehead wie folgt kommentiert: "From this point of view, paradoxically enough, logic is not concerned with developing our powers of thought but with developing techniques that enable us to get along without thinking!" (p. 213).

Augenzeugen, so dass von vorneherein Vergleiche zwischen verschiedenen Beobachtungsberichten ausgeschlossen sind. Außerdem entziehen sich viele relevante Prozesse der unmittelbaren Anschauung: man sieht einer elektronischen Bremsanlage nicht an, ob sie defekt ist. In der Naturwissenschaft besteht hingegen die Möglichkeit der kontrollierten, wiederholten und mit unterschiedlichen Methoden durchgeführten Beobachtung und Messung. Dazu wurden normative Standards entwickelt, die die Intersubjektivität, Objektivität und Präzision der Beobachtung garantieren.

**Intersubjektivität** erfordert, dass Beobachtungen und experimentelle Befunde prinzipiell für jedes kompetente Mitglied der Wissenschaftsgemeinschaft nachvollziehbar sein sollen. Damit mehrere Personen ihre Beobachtungen vergleichen können, müssen diese Beobachtungen möglichst fehlerfrei kommuniziert werden können. So trivial diese Forderung ist, so schwierig ist sie ohne fehlerfrei kopierbare Medien zu erfüllen. Mündliche Berichte scheiden von vorneherein aus, aber auch zeichnerische Darstellungen konnten vor der Erfindung von Buchdruck, Holzschnitt- und Gravurtechniken praktisch nicht fehlerfrei vervielfältigt werden. Digitale Worte und Ziffern unterscheiden sich in dieser Hinsicht grundlegend von analogen Zeichnungen und Diagrammen. In antiken und mittelalterliche Abschriften von Reiseberichten, Zahlenwerken etc. haben sich natürlich auch zahlreiche Fehler eingeschlichen, aber im Gegensatz zu den Fehlern, die sich beim Abzeichnen z.B. von Karten ergaben, sind diese zu vernachlässigen. Sie wurden deshalb auch nicht in nennenswertem Umfang zur Darstellung von räumlichen Daten und technischen Geräten verwendet (Ivins 1969). Während Karten und technische Zeichnungen sich nach der Erfindung der Druck- und Gravurtechniken aufgrund ihrer ökonomischen Relevanz sehr schnell durchsetzten, wurden rein wissenschaftliche nicht-räumliche Beobachtungswerte (Luftdruck, Temperaturen, etc.) noch lange Zeit primär in numerischen Tabellen festgehalten und disseminiert.[7]

Vor diesem Hintergrund wird deutlich, auf welch abstraktem und extrem kondensiert Informationsangebot die neuzeitliche Mathematisierung der Naturwissenschaften beruht. Zwar spielten anschauliche geometrische Beweisformen auch lange nach Erfindung der analytischen Geometrie noch eine wichtige Rolle, aber diese geometrischen Figuren dienten einzig der mathematischen Beweisführung und nicht der konkreten Darstellung von beobachteten Ereignissen. D.h. die Idee, dass im Gegensatz zum aristotelischen Weltbild alle sub- und supralunaren Erscheinungen uniformen und allgemeinen quantitativen Naturgesetzen gehorchen, setzten sich zu einer Zeit durch, als die verfügbaren Medien für Beobachtungen statisch, symbolisch und abstrakt waren. Die anschaulich erlebbare Dynamik von Billiardstößen wurden ebenso wie die nicht direkt wahrnehmbaren Planetenbewegungen auf Messwerte reduziert. Die intersubjektive Bestätigung der neuen quantitativen Gesetze lag dementsprechend primär darin, dass numerische Daten mit theoretische Berechnungen auf dem Papier verglichen wurden. Der direkte anschauliche Zugang zu Naturphänomenen wurde zwar im Experiment kultiviert, wurde aber erst durch die Verwendung von Medien überprüfbar. Nicht die Phänomene selbst, sondern mathematische Darstellungen auf dem Papier waren die eigentlichen Gegenstände der wissenschaftlichen Reflexion. Das subjektive unmittelbare Erleben im Sinne der phänomenalen Kausalität war dadurch bei der Begründung kausalen Wissens weitgehend ausgeblendet worden.

**Objektivität** der Beobachtung fordert, dass selbst psychische Einflußfaktoren, die allen Menschen gemeinsam sind, wie z.B. Ermüdungserscheinungen, minimiert werden. Im Idealfall tritt deshalb an die Stelle des menschlichen kognitiven Systems ein Automat, der

[7] Erst Johann Heinrich Lambert verwendete im 18. Jahrhundert Graphen zur Veranschaulichung von Messwerten. Durchgesetzt haben sich diese graphische Visualisierungen von Daten jedoch erst im neunzehnten Jahrhundert, obwohl mit der analytischen Geometrie von Descartes und Fermat die mathematischen Grundlagen dafür schon lange bereit standen (Tilling 1975).

völlig frei von menschlichen Urteilsverzerrungen ist. Solche "reinen" Beobachtungen ohne beobachtendes Subjekt liegen dann vor, wenn der Messapparat direkt numerische Werte ausgibt, die ohne weitere Übersetzungs- und Deutungsschritte in Berechnungen eingehen können.[8] Ein Messgerät, das auf seinem Display "Kausalgrade" anzeigt, ist hingegen meines Wissens bislang nicht erfunden worden. Allein aus diesem Grund sind Kausalbeziehungen nicht messbar und damit auch nicht objektiv beobachtbar.

Dabei sollte aber nicht vergessen werden, dass die Physik heute über eine differenzierte Begrifflichkeit verfügt, bei der Eigenschaften als messbar gelten, die man als kausal bezeichnen könnte. Hume beklagte (wohl zu Recht), dass viele kausale Begriffe nahezu synonym verwendet werden: "efficacy, agency, power, force, energy, necessity, connection, and productive quality, are all nearly synonimous; and therefore 'tis an absurdity to employ any of them in defining the rest." (Hume 1978, S. 157). Einige dieser Ausdrücke haben in der heutigen Physik allgemein akzeptierte Definitionen. Wissenschaftsgeschichtlich könnte man diese Entwicklung als Ausdifferenzierung des alltagssprachlichen Ursachebegriffs in exakt definerte quantitative Impuls-, Kraft-, Arbeits-, Leistungs- und Energiebegriffe beschreiben. Auch diese Entwicklung setzte erst ein, nachdem zunehmend Formeln die normalsprachlichen wissenschaftlichen Texte ergänzten.[9] Bemerkenswert daran ist, dass einige der von Hume angeführten Begriffe (wie z.B. Kraft und Energie, die man umgangsprachlich vielleicht als "Veränderungspotential" oder "mögliche Wirksamkeit" bezeichnen könnte) in der modernen Physik durchaus als messbar angesehen werden. Was also als objektiv messbar gilt, ist relativ zur instrumentellen Ausstattung und den zugrunde liegenden Operationalisierungsvorschriften. Ein für alle Zeiten feststehender Konsens hinsichtlich der Beobachtbarkeit von Qualitäten ist deshalb nicht zu erwarten.

**Präzision.** Obwohl die Bedeutung von Messgeräten auch im Alltag zunimmt, dürfte dies bislang wohl kaum einen Einfluss auf das Alltagsverständnis von Kausalität gehabt haben. Dies war in der neuzeitlichen Physik anderes, in der immer wieder debattiert wurde, ob es eine obere Grenze für Wirkungsausbreitungen gibt oder nicht. Descartes war z.B. der Meinung, dass es zwar keine Fernwirkungen gibt, sich mechanische Stöße aber über mehrere Vermittlungstationen und verschiedene Raumstellen hinweg instantan auswirken können. Die Newtonsche Physik lies mit ihrem Gravitationsgesetz sowohl Fern- als auch Instantanwirkungen zu, während die Einsteinsche Relativitätstheorie Fern- und Instantwirkungen ausschließt und die Lichtgeschwindigkeit als obere Grenze der Wirkungsausbreitung ansieht. Die Frage, welche dieser Auffassungen richtig ist, ist offensichtlich mit dem bloßen Auge nicht zu entscheiden, obwohl sie für die philosophische Frage, ob die Asymmetrie der Kausalrelation auf die Asymmetrie der Zeit zurückgeführt werden kann, von ganz entscheidender Bedeutung ist. So setzt die von Hume postulierte Beobachtbarkeit der zeitlichen Priorität der Ursache vor der Wirkung, in vielen Fällen genauere Zeitmessungen voraus, als sie im Alltag möglich sind. Wenn jemand ein Glas umgreift, es festhält und plötzlich verschiebt, ist jedenfalls nur eine Gleichzeitigkeit der Bewegungen der Hand und des Glases wahrnehmbar, obwohl natürlich physikalisch gesehen die Wände des Glases minimal eingedrückt werden, bevor das Glas als Ganzes sich in Bewegung setzt. Oder anders ausgedrückt: Die bei solchen Fällen für selbstverständlich gehaltene

---

[8] Erste automatische Aufzeichnungssysteme datieren auf die zweite Hälfte des 17. Jahrhunderts. 1663 erfand z.B. Wren eine Wetteruhr, ein Uhrwerk getriebenes Aufzeichnungsgerät für Windrichtung, Regenmenge und Temperatur.

[9] Zur Erläuterung der Definitionen werden von Wissenschaftshistorikern oft Formeln angeführt, die so meist gar nicht von den entsprechenden Autoren benutzt wurden: Der Kraftbegriff Descartes wird z.B. als $m \cdot v$, der von Leibniz als $m \cdot v^2$ und der von Newton als $m \cdot a$ definiert. Dahinter steckt die Auffassung, dass erst mit der Formel eine klare Definition eines zentralen Begriffs der Physik gelungen sei (wie z.B. die Definition der Kraft in Newtons Axiomen).

Asymmetrie von Ursache und Wirkung kann nicht auf wahrnehmbaren zeitlichen Unterschieden sondern muss auf Vorwissen beruhen, dass wiederum erst mit der Mathematisierung der Physik und der Entwicklung von Präzisionsinstrumenten gewonnen werden konnte.

**Fazit**

Mit der Mathematisierung der Wissenschaften wurde eine Mischung aus Normalsprache und mathematischer Idealsprache zum Standard der Wissenschaftlichkeit. Damit wurden Mythen und andere narrative Formen der Kausalerklärung nach und nach durch allgemeine und kohärente wissenschaftliche Argumentationen ersetzt (Donald 1991; Lindberg 1992), in der Beobachtungen und deren Interpretation systematisch unterschieden werden können. Diese langfristige und keineswegs geradlinige Entwicklung hatte in der Physik mit Isaac Newton einen ersten Höhepunkt erreicht, womit ein auch heute noch geltendes Ideal für andere Wissenschaften vorgezeichnet war.

Hume war vielleicht nicht im Einzelnen mit den Postulaten der neuzeitlichen Physik vertraut (Hume bestritt z.B. Instantan- und Fernwirkungen) aber er war sich der generellen Bedeutung der mathematischen Physik sicherlich bewußt. Die humesche Grundidee, dass das spezifisch kausale an Ereignissequenzen letztlich in nichts anderem als der Instanziierung von allgemeinen Gesetzen besteht, bezieht ihre große Akzeptanz sicherlich auch aus den Erfolgen der klassischen Physik, der es gelang, so phänomenal disparate Vorgänge wie Planetenbewegungen, Gezeiten und Billiardstöße unter wenige quantitative Gesetze zu subsumieren. Vor diesem Hintergrund läßt sich auch die Attraktivität der Humeschen Regularitätstheorie der Kausalität erklären.

Die kognitiven Voraussetzungen dafür werden jedoch von Kausalitätstheoretikern in der Regel nicht thematisiert. Weder Hume noch seine Nachfolger in Wissenschaftstheorie und Psychologie können plausibel erklären, wie Menschen überhaupt zu dem abstrakten kausalen Wissen kommen sollen, dass z.B. Naturgesetze im sub- und supralunaren Bereich gleichermaßen gelten, Ursachen immer zeitlich vor der Wirkung auftreten und Fernwirkungen auszuschließen sind. Solch abstraktes kausales Wissen geht zwar in moderne Definitionen von Kausalität mit ein, es kann aber nicht der persönlichen Erfahrung und Anschauung entstammen. Die Anforderung die die Humesche Regularitätstheorie und ihre modernen Varianten an das kognitive System stellen, sind ohne Hilfsmittel praktisch nicht zu erfüllen: Bei der korrekten Analyse der Beobachtbarkeit von Kausalbeziehungen soll vorgängiges Kausalwissen vollständig ausgeblendet werden, obwohl auch nach Humeschen Verständnis dieses kausale Vorwissen im Alltag unwillkürlich mit in die Interpretation von Beobachtung eingeht. Quantitative nicht-natürlichsprachliche Notationen helfen dem kognitiven System, vom kausalen Gehalt der Alltagssprache und Alltagsbeobachtungen zu abstrahieren. Dabei sollte allerdings nicht vergessen werden, wie schwer den meisten Menschen der Umgang mit formalen Notationen fällt.

## Literatur


Copi, I. M. (1968). *Introduction to Logic* (3rd ed.). London: Macmillan 1968.

Ducasse, C. J. (1993) On the Nature and the Observability of the Causal Relation. In: E. Sosa, & M. Tooley (eds.), *Causation*, (pp. 125-136). Oxford: Oxford University Press.

Fiedler, K., Semin, G. R., & Finkenauer, C. (1994) Welchen Spielraum läßt die Sprache für die Attribution. In: F. Försterling, & J. Stiensmeier-Pelster (eds.), *Attributionstheorie*, (pp. 27-54). Göttingen: Hofgrebe.

Gigerenzer, G. (1987). Probabilistic Thinking and the fight against Subjectivity. In: L. Krüger, G. Gigerenzer, M.S. Morgan (eds.), *The Probabilistic Revolution, Vol. II: Ideas in the Sciences, (pp. 11-34).* Cambridge, MA: MIT Press.

Switjink, Z. G. (1987) The Objectification of Observation: Measurement and Statistical Methods in the Nineteenth Century. In: L. Krüger, L. J. Daston, M. Heidelberger (eds.), *The Probabilistic Revolution, Vol. I: Ideas in History, (pp. 261-285).* Cambridge, MA: MIT Press.

Goody, J. (1987). *The interface between the written and the oral.* Cambridge: Cambridge University Press.

Hume, D. (1978) *A Treatise of Human Nature.* L. A. Selby-Bigge (ed.), (2nd ed.), Oxford: Clarendon Press.

Hutchins, E. (1995). *Cognition in the Wild.* Cambridge, MA: MIT Press.

Ivins, William M. (1969/1953). *Prints and Visual Communication.* Cambridge, Mass.: MIT Press.

Kelsen, H. (1941). *Vergeltung und Kausalität.* The Hague, Holland: W. P. Van Stockum & Zoon.

Lindberg, D. C. (1992). *The Beginnings of Western Science.* Chicago: University of Chicago Press.

Mackie, J. L. (1974) *The Cement of the Universe*. Oxford: Oxford University Press.

Menninger, K. (1969). *Number words and number symbols: a cultural history of numbers*. Cambridge, MA: MIT Press.

Michotte, A. (1982) *Gesammelte Werke*. O. Heller (ed.), Vol. 1, Bern: Huber.

Olson, D. R. (1996). Towards a psychology of literacy: on the relations between speech and writing. *Cognition, 60*, 83-104.

Rock, I. (1983). *The logic of perception.* Cambridge, MA: MIT Press.

Ross, M., & Fletcher, G. J. O. (1985) Attribution and Social Perception. In: G. Lindzey, & Aronson E. (eds.), *The Handbook of Social Psychology* (3rd ed.), (pp. 73-122). New York: Random House.

Rumelhart, D. E. (1989). Towards a Microstructural Account of Human Reasoning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 298 - 312). Cambridge: Cambridge University Press.

Stegmüller, W. (1983) *Probleme und Resultate der Wissenschaftstheorie und analytischen Philosophie, Bd. I*. Berlin: Springer.

Strawson, P. F. (1985) Causation and Explanation. In B. Vermazen, & M. B. Hintikka (Eds.), *Essays on Davidson - Action and Events*, (pp. 115-135). Oxford: Clarendon.

Tilling, L. (1975). Early experimental graphs. *The British Journal for the History of Science*, Vol. 8, No. 30, 193-213.

Warren, W. H., Nicholas, D. W., & Trabasso, T. (1979). Event chains and inferences in understanding narratives. In R. O. Freedle (Ed.) *New directions in discourse processing. Vol. II* (pp. 23-52). Norwood: Ablex.

Weinberg, J. (1973) "Causation". In P. P. Wiener (ed.), *Dictionary of the history of ideas*, Vol. I, (pp. 270-278).

# Die Kontextabhängigkeit kausaler Erklärungen

Michael May
GMD – Forschungszentrum Informationstechnik,
Schloß Birlinghoven, Sankt Augustin,
michael.may@gmd.de

## 1 Einleitung

Der Fernsehreporter steht vor den rauchenden Trümmern des Flugzeugs. Rauch brennt ihm in den Augen. Zu den Zuschauern gewandt fragt er: „Warum ist das Flugzeug abgestürzt?“. Er *weiß*, daß es abgestürzt ist. Was er zusammen mit den Zuschauern verlangt, ist eine *Erklärung*.

Es gibt eine Vielzahl von Ansätzen, den Begriff der Erklärung formal zu fassen. Trotz ihrer sonstigen Heterogenität stimmen sie überein, daß von dem Wissen um das zu erklärende Phänomen in der Erklärung kein Gebrauch gemacht werden darf. Das ist überraschend, denn es liegt nahe, dieses Wissen in der Beantwortung der Frage des Reporters zu verwenden. In den meisten Erklärungsansätzen führt dies jedoch entweder zu einer Trivialisierung der Erklärung oder sie wird inkorrekt. Die wichtigsten formalen Erklärungstheorien sollen im Hinblick auf dieses Phänomen untersucht werden. Anschließend werde ich einige Schwierigkeiten aufzeigen, die sich hieraus ergeben. Schließlich werde ich einen Ansatz zur kausalen Erklärung skizzieren, der es erlaubt, daß Wissen um das zu erklärende Phänomen vorauszusetzen und dadurch zu einem viel natürlicheren Modell führt, welches nebenbei erlaubt, der Kontextabhängigkeit von Erklärungen Rechnung zu tragen.

## 2 Modelle der Erklärung

### 2.1 Erklärung als Deduktion

Die Auffassung, daß eine Erklärung die Form eines gültigen deduktiven Argumentes hat, wobei der Beweis die Erklärung und das zu erklärende Phänomen das Theorem ist, wurde zuerst von Hempel (1948, 1965) im deduktiv-nomologischen Erklärungsmodell formal ausgearbeitet. Gegeben eine Menge von empirischen Fakten $A_1, ..., A_n$ und eine Menge von Gesetzesaussagen $L_1, ..., L_m$. Die Grundidee ist, daß die $A_i$ und $L_i$ eine Erklärung für $E$ sind, wenn $E$ aus ihnen logisch ableitbar ist:

$$A_1, ..., A_n, L_1, ..., L_m \vdash E,$$

wobei mindestens ein Gesetz $L_i$ wesentlich in dem Argument vorkommt, d.h. $A_1, ..., A_n, \{L_1, ..., L_m\} - \{L_i\} \nvdash E$. Ein deduktiver Erklärungsbegriff findet sich im Anschluß an Peirce häufig in Theorien der Abduktion und des Schlußfolgerns auf die beste Erklärung. Auch im Maschinellen Lernen, etwa dem erklärungsbasierten Lernen (Mitchell 1996; S. 313ff) und der Induktiven Logikprogrammierung (z.B. Srinivasan 1999) ist diese Auffassung weit verbreitet. Da das zu erklärende Phänomen $E$ das abzuleitende Theorem ist, darf es nicht unter den Prämissen vorkommen, da sich der Beweis dann sofort trivialisiert und ein Ereignis durch sich selbst „erklärt“ würde.

### 2.2 Erklärung als induktives Argument

Ein ebenso offensichtliches wie schwerwiegendes Problem der deduktiven Auffassung der Erklärung ist, daß die geforderten deduktiven Argumente praktisch nie bekannt sind. Dadurch wird die praktische Anwendbarkeit des Modells untergraben. Dies läßt die Idee aufkommen, die Implikationsrelation abzuschwächen. Das kann in verschiedenen Formen geschehen. Ebenfalls von Hempel stammt das induktiv-statistische Modell (Hempel 1965). Es hat als wesentliche Elemente sta-

tistische Gesetze $L_1, ..., L_m$ sowie empirische Fakten $A_1, ..., A_n$, aus denen mit hoher Wahrscheinlichkeit folgt, daß das zu erklärende Phänomen $E$ eintritt. Von der Intention verwandt sind Ansätze, welche eine nicht-monotone Logik zugrundelegen (z.B. Geffner 1992, Boutilier & Becher 1995). In allen genannten Ansätzen ist klar, daß das Wissen um $E$ nicht vorausgesetzt werden darf. Wird es vorausgesetzt, verliert das Argument erstens seinen induktiven Charakter und wird zweitens trivial, da wiederum $E$ durch sich selbst erklärt würde.

Den induktiven und deduktiven Argumenten ist gemeinsam, daß sie eine Erklärung in Analogie zu einer – deterministischen, statistischen oder „plausiblen" – Prognose des zu erklärenden Ereignisses auffassen.

## 2.3 Erklärung als statistische Relevanz

Die Begründung des Erklärungsbegriffs über die statistische Relevanz wurde u.a. von Salmon (1971, 1984 u.a.) ausgearbeitet und enthält als ein wesentliches Element die statistische Ungleichung

$$P(E|C) \neq P(E).$$

Eine Erklärung $C$ zeichnet sich dadurch aus, daß sie statistisch relevant für das zu erklärende Phänomen $E$ ist und normalerweise die Wahrscheinlichkeit des Eintretens des zu erklärenden Phänomens erhöht (Salmon selbst erlaubt, daß die Erklärung negativ relevant ist; darin jedoch sind ihm viele Autoren, etwa Gärdenfors (1988), nicht gefolgt). Wird $E$ als gegeben vorausgesetzt, so ist $P(E) = P(E|C) = 1$ und die Ungleichung damit falsch. Die statistische Relevanz ist auch in dem pragmatischen Ansatz van Fraassens (1980) ein zentrales Element.

## 2.4 Maximum a posteriori Wahrscheinlichkeit (MAP)

Ein anderer Ansatz, der in der Bayesschen Statistik oft verwendet wird, ist nicht prognostisch, sondern *diagnostisch* orientiert (Pearl 1988). Gesucht wird die *aposteriori* wahrscheinlichste Ursache des zu erklärenden Phänomens, d.h. dasjenige $C$, welches durch die Evidenz $E$ maximiert wird. Auf den ersten Blick kann hier das zu erklärende Phänomen vorausgesetzt werden, da nicht die Wirkung (das zu erklärende Phänomen selbst) sondern die Ursache erschlossen werden soll. Die Berechnung erfolgt über das Bayessche Theorem, indem diejenige Ursache mit der höchsten *a posteriori* Wahrscheinlichkeit als die beste (=wahrscheinlichste) Erklärung ausgezeichnet wird:

$$\max_{C_i \in \{C_i\}} P(C_i|E) = \frac{P(E|C_i)\ P(C_i)}{P(E)}.$$

Chajewska & Halpern (1997) weisen in einer Kritik von Pearls Ansatz darauf hin, daß sich hier dasselbe Problem wie in den zuvor diskutierten Ansätzen stellt: Wird $E$ als gegeben vorausgesetzt, so ist $P(E) = P(E|C_i) = 1$ und das Problem reduziert sich auf

$$\max_{C_i \in \{C_i\}} P(C_i),$$

d.h. es wird die Hypothese mit der größten *apriori*-Wahrscheinlichkeit als Erklärung ausgewählt. Die Bayessche Methodologie ist darauf ausgelegt, die Änderung der Wahrscheinlichkeit aufgrund zusätzlicher Evidenz zu berechnen und hilft nicht bei der Berechnung der apriori-Wahrscheinlichkeit. Damit erweist er sich bei bereits gegebenem $E$ als überflüssig. Um einer derartigen Konsequenz zu entgehen, argumentieren Howson & Urbach (1993, S. 404f), daß jemand, der im Bayesschen Modell $P(E) = 1$ setzt, das Modell inkorrekt verwendet und eine korrekte Verwendung die Wahrheit von $E$ nicht voraussetzen darf. Daraus folgt aber ebenfalls, was zu zeigen war: daß sich auch das diagnostische Bayessche Modell – entgegen dem ersten Anschein – in dieser Frage nicht von den anderen Modellen unterscheidet.

## 2.5 Epistemische Relativierung

Wir haben gesehen, daß die dargestellten Modelle die Eigenschaft teilen, daß von dem Wissen um das zu erklärende Phänomen in der Erklärung kein Gebrauch gemacht werden darf. Nun soll gezeigt werden, daß sich daraus Schwierigkeiten ergeben.

Von den dargestellten Modellen enthalten einige, insbesondere das induktiv-statistische und das Bayessche Modell, eine Relativierung auf eine Wissenssituation. Eine solche epistemische Relativierung ist aus einer kognitiven Perspektive heraus wünschenswert, weil dann das Wissen von epistemischen Subjekten modelliert werden kann. Aber

auch aus systematischen Gründen muß eine geeignete Relativierung vorgenommen werden, wie am Beispiel des statistischen Relevanzmodells gezeigt werden soll. Der Grund ist ein Phänomen, das unter Namen wie *Problem der Referenzklasse* oder *Simpsons Paradox* bekannt ist, und welches sich aus dem nicht-monotonen Charakter der Statistik herleitet: wenn relativ zu einer Bezugsklasse $K$ die Relation

$$P(E|C) > P(E)$$

gilt, dann kann in einer Subklasse von $K$ die Negation dieser Aussage wahr sein; relativ zu $K' \subset K$ kann gelten

$$P(E|C) \leq P(E).$$

Beispielsweise mag die Wahrscheinlichkeit für Menschen, den Sprung von einem Haus unverletzt zu überstehen, kleiner sein als die Wahrscheinlichkeit, sich zu verletzen. Für eine Teilklasse der Sprünge, nämlich denen von professionellen Stuntmen, mag die Wahrscheinlichkeit unverletzt zu bleiben, hingegen größer sein. Relativiert man hingegen weiter auf die Sprünge von Stuntmen von extrem hohen Häusern, so mag die Wahrscheinlichkeit unverletzt zu bleiben wiederum kleiner sein usw. Wird daher nicht spezifiziert, auf welche Referenzklasse relativiert werden soll, dann erscheint es willkürlich, ob $C$ eine Erklärung für $E$ ist oder aber nicht, je nachdem auf welche Teilmenge des Wissens einer Person relativiert wird.[1]

Sowohl aus kognitiver als auch aus systematischer Sicht ist es naheliegend, die Wahrscheinlichkeitsaussage auf das *gesamte* gegenwärtige Wissen des Fragenden zu relativieren. Eine solche Forderung kommt beispielsweise in Carnaps Forderung nach der Berücksichtigung der vollständigen Evidenz (Carnap 1950) und Reichenbachs Prinzip der engsten Bezugsklasse zum Ausdruck (Reichenbach 1949). *Dieser Weg ist jedoch versperrt.* Das Wissen der Person umfaßt normalerweise, wie anfangs festgestellt wurde, das zu erklärende Phänomen $E$, und kann deshalb – aus den dargestellten Gründen – nicht als Referenzklasse genommen werden. Wie kann nun eine Relativierung erreicht werden, ohne daß es zu einer trivialen oder inkorrekten Erklärung kommt? Es erscheint klar, daß irgendeine Teilmenge des gegenwärtigen Wissens genommen werde muß. Aber wie ist sie bestimmt?

[1] Für eine aktuelle Darstellung zu Simpsons Paradox siehe Pearl (1999). Für dieses Phänomen gibt es im deduktiven Erklärungsmodell keine Entsprechung: In der deduktiven Logik, die monoton ist, gilt: wenn $A, L \vdash E$ dann auch $A, L, X \vdash E$ , für beliebiges $X$. Die Gültigkeit eines deduktiven Argumentes kann durch zusätzliche Evidenz nicht untergraben werden.

## 2.6 Epistemische Relativierung und Überzeugungsrevision

Ein interessanter Ansatz stammt von Peter Gärdenfors (1988). Er kombiniert seine Theorie der Überzeugungssysteme (*belief systems*) mit dem statistischen Relevanzmodell. Ein Überzeugungssystem ist eine konsistente, unter der logischen Implikation abgeschlossene Menge von Aussagen. Wenn $K$ das gegenwärtige Wissen des Erklärungssuchenden ist, welches normalerweise $E$ enthält, dann ist $K_E^-$ das um die Überzeugung $E$ kontrahierte Überzeugungssystem. Eine Kontraktion ist eine logische Operation, durch die eine akzeptierte Aussage aus einem Überzeugungssystem entfernt wird, so daß das kontrahierte System bestimmten Bedingungen genügt – grob gesagt, daß es die epistemisch am besten verankerte maximale Teilmenge des ursprünglichen Systems ist, aus der die entfernte Aussage nicht ableitbar ist. Damit $C$ eine Erklärung für $E$ ist, muß gelten

$$P(E|C.K_E^-) > P(E|K_E^-).$$

Gärdenfors erhebt zusätzlich die Forderung der vollständigen Evidenz bezüglich $K_E^-$, daß es also das *gesamte* für $E$ relevante Wissen des Fragenden enthält, das nach der Kontraktion in $K_E^-$ verbleibt. Wenn der Fragende sich im epistemischen Zustand $K$ befindet, welcher $E$ umfaßt, dann wird also zunächst durch eine Kontraktion ein Zustand $K_E^-$ gebildet, in dem $E$ nicht enthalten bzw. ableitbar ist; dann werden Aussagen gesucht, deren Hinzufügen zu $K_E^-$ die Wahrscheinlichkeit von $E$ erhöht. Indem diese Aussagen $E$ wahrscheinlicher machen, vermindern sie die „kognitive Dissonanz“ (Gärdenfors) des Fragenden.

Die Kontraktion ist keine simple Operation, da nicht nur $E$ entfernt werden muß, sondern auch alle Aussagen, aus denen $E$ ableitbar ist. Es ist wohlbekannt, daß es im allgemeinen verschiedene Wege gibt, eine solche Kontraktion zu erreichen. Soll z.B. das System

$$K = \{p, p \rightarrow q, q\}$$

um $q$ revidiert werden, so reicht es nicht, $p$ herauszunehmen, da $p \wedge (p \rightarrow q) \vdash q$; es kann aber z.B. durch Kontraktion auf

$$K_E^- = \{p\}$$

oder auch

$$K_E^- = \{p \rightarrow q\}$$

geschehen, da sowohl $p \nvdash q$ als auch $(p \rightarrow q) \nvdash q$ gilt. Diese beiden Systeme sind aber nicht äquivalent.

Eine weitere Schwierigkeit ist, daß für hinreichend mächtige Sprachen kein Entscheidungsverfahren existiert, welches im allgemeinen Fall angibt, ob eine Aussage ableitbar ist oder nicht. Bereits die Prädikatenlogik 1. Stufe ist ein solcher Fall: Es existiert keine allgemeine Methode, um Allgemeingültigkeit zu widerlegen. Wir sehen also, daß für epistemisch relativierte Theorien die Forderung, $E$ nicht zu benutzen, keine trivial zu erfüllende Bedingung ist, sondern technisch sehr aufwendige Konsequenzen hat.

Ein weiteres Problem ist, daß weder $K_E^-$ noch das um die erklärungsrelevanten Fakten expandierte System ein Überzeugungssystem ist, in dem sich der Fragende befindet oder auch nur befunden haben muß. Es erscheint jedoch für einen pragmatisch und kognitiv motivierten Ansatz enorm unplausibel, die Verminderung der kognitiven Dissonanz in Bezug auf zwei Überzeugungssysteme zu bestimmen, *von denen weder das eine noch das andere dasjenige des Fragenden ist.*

Sowohl Chajewska & Halpern (1997) als auch Howson & Urbach (1993, S. 406) schränken ihre Ansätze angesichts der Schwierigkeiten einer eindeutigen Relativierung auf den einfachen Fall ein, wo $E$ logisch unabhängig von den übrigen Aussagen in $K$ ist, so daß

$$K_E^- = K - \{E\},$$

d.h. durch mengentheoretische Subtraktion gewonnen werden kann. Damit wird jedoch die Reichweite ihres Modells extrem eingeschränkt und ist auf viele interessante Fälle nicht mehr anwendbar. Gehen wir noch einmal zurück zu dem Reporter. In seinem gegenwärtigen Überzeugungssystem nimmt das Wissen um den Absturz des Flugzeuges eine zentrale Rolle ein: es ist der Grund dafür, daß er selbst an diesen Ort geschickt wurde, um eine Reportage zu machen; es erklärt die Anwesenheit anderer Reporter, der Löschfahrzeuge, der Krankenwagen. Es erklärt die Verspätung am Flughafen; es bestimmt möglicherweise seine Absichten und Handlungen für die nächsten Tage. Sein Wissen um den Absturz ist sicherlich nicht unabhängig von seinem übrigen Wissen, welches auch das Wissen um die Wirkungen des Absturzes umfaßt. Wir müssen uns jedoch auf ein Überzeugungssystem des Reporters beziehen, aus dem nicht unmittelbar ableitbar ist, daß das Flugzeug abgestürzt ist. In Fällen wie dem dargestellten setzt das aber voraus, daß wir wesentliche Teile des gerade jetzt relevanten Wissens wegschneiden. Wie wir noch sehen werden, ist das Wissen um $E$ eine wertvolle Hilfe bei der Konstruktion einer Erklärung, auf das wir hier jedoch verzichten müssen.

Ich habe argumentiert, daß die betrachteten Ansätze a) das Wissen um $E$ in einer Erklärung nicht verwenden können und b) dadurch in Schwierigkeiten bei der notwendigen epistemischen Relativierung geraten. Die statistischen Modelle gewinnen auf den ersten Blick an Realitätsnähe und psychologischer Plausibilität, indem sie Erklärung als kontextgebundenes Phänomen analysieren, welches auf eine epistemische Situation relativiert wird. Dieser Vorteil wird aber sofort in Frage gestellt dadurch, daß auf eine *kontrafaktische* Situation relativiert wird, mithin also nicht auf den epistemischen Kontext, in dem die Person sich tatsächlich gerade befindet. Das deduktiv-nomologische Modell vermeidet zwar das Problem der epistemischen Relativierung, aber um einen zu hohen Preis, da solche deduktiven Argumente normalerweise nicht zur Verfügung stehen.

# 3 Erklärung durch Elimination von Alternativen

Im folgenden wird ein Erklärungsmodell skizziert, welches die Relativierung auf eine epistemische Situation erlaubt, in der $E$ bereits gewußt wird. Das Modell hat eine Reihe positiver Eigenschaften. Insbesondere erlaubt es die Relativierung auf das tatsächliches Wissen der Person, wodurch die Evidenz $E$ benutzt und das Problem der Referenzklasse vermieden werden kann. Weiterhin er-

laubt es auf natürliche Weise, Kontextphänomene zu berücksichtigen. Das Modell ist durch drei wesentlich Elemente bestimmt.

**1. Ein Ereignis zu erklären heißt, seine Ursache zu identifizieren.**

Die Erklärung für $E$ ist in diesem Modell einfach die Ursache von $E$. Wenn der Reporter sagt, daß wir die Erklärung noch nicht gefunden haben, meint er nach diesem Modell, daß die *Ursachen* noch nicht gefunden worden sind. Das Modell ist ein kausales diagnostisches Modell der Erklärung. Ziel der Erklärung ist es nicht, Evidenzen für $E$ herbeizuschaffen, von dessen Bestehen der Fragende ohnehin überzeugt ist, sondern Evidenzen dafür, daß eine der möglichen Ursachen $C_1, ..., C_n$ im gegenwärtigen Kontext die tatsächlich wirksame Ursache war.

**2. Eine Ursache wird identifiziert, indem alternative Ursachen eliminiert werden.**

Die Identifikation der erklärenden Ursache geschieht durch einen *eliminativen Prozess*:

1. Gegeben eine Menge $\mathcal{M}$ von möglichen Ursachen
2. Schließe alle Ursachen bis auf eine aus
3. Die verbleibende Ursache ist die Erklärung des Phänomens

Angenommen, es gibt im gegenwärtigen Kontext nur vier alternative potentielle Ursachen, die für den Flugzeugabsturz in Betracht kommen: ein Triebwerkschaden, eine Bombe im Frachtraum, ein Bedienfehler des Piloten, ein Unwetter. Die Untersuchung der Triebwerke zeigt, daß sie in Ordnung waren. Die Lage und Gestalt der Trümmer zeigt, daß eine Bombenexplosion auszuschließen ist. Auch ein Unwetter kann aufgrund der Gesamtwetterlage sicher ausgeschlossen werden. Aber da Flugzeuge nicht einfach so vom Himmel fallen, muß etwas das Unglück verursacht haben. Die Auswertung des aufgezeichneten Gespräche im Cockpit deutet auf einen Bedienfehler des Piloten hin, der aber normalerweise harmlos ist. Unter der Annahme, daß die betrachteten Ursachen die einzigen sind, die in Betracht kommen, und der Bedienfehler die einzig verbleibende Möglichkeit ist, ist dies unsere beste Erklärung. Aus einer Menge von Ursachen werden also solange Alternativen entfernt, bis nur noch eine übrig bleibt.

Die Falschheit einer Aussage läßt sich oft einfacher nachweisen läßt als ihre Wahrheit – das ist die Grundmotivation des eliminativen Schließens. Zu zeigen, daß etwas nicht die Ursache sein kann, ist oft einfach; auf direktem Wege zu zeigen, daß es die Ursache war, hingegen oft praktisch unmöglich. Bedienfehler führen keineswegs *immer* zum Absturz; es kann aufgrund der eingebauten Sicherheitsmaßnahmen sogar sehr unwahrscheinlich sein. Die Wahrscheinlichkeit der Wirkung, gegeben eine der Ursachen, ist für das Argument unerheblich; sie kann beliebig niedrig sein. Dennoch können wir darauf schließen, daß der Bedienfehler ein relevanter Teil der Ursache war, da dies im gegenwärtigen Kontext die *einzige verbleibende* Erklärung ist. Dies ist eine direkte Anwendung des berühmten Sherlock-Holmes-Prinzips: „Whenever you have eliminated the impossible, whatever remains, however improbable, must be the truth.“

Der Prozeß der Identifikation der Ursachen kann für einige Zwecke in die Form eines Argumentes gebracht werden. Wenn $C_1$, $C_2$, ..., $C_n$ die vollständige Menge der potentiellen Erklärungen ist, ist dies ein deduktiv gültiger Schluß:

Nur wenn $C_1$ oder $C_2$ oder ... oder $C_n$, dann $E$
$E$ ist der Fall
$C_1$, $C_2$, ..., $C_{n-1}$ sind nicht der Fall
---
$C_n$

Das Wissen um $E$ spielt in diesem Argument eine wesentliche Rolle. Ohne $E$ ließe sich die Elimination nicht durchführen, da nur das vorausgesetzte Wissen um $E$ uns zu rechtfertigen erlaubt, daß eines der $C_i$ gegeben sein muß.

**3. Die Ursache erklärt, weshalb die Wirkung im zu erklärenden Fall eingetreten ist, in vergleichbaren, durch den Fragekontext implizierten Fällen hingegen nicht.**

Eine wesentliche Frage ist, wie die Menge der möglichen Ursachen in einem gegebenen Kontext auf eine kleine Anzahl eingegrenzt werden kann. Damit sich ein eliminativer Schluß in praktischen

Kontexten auf eine überschaubare Menge von Alternativen stützen kann, ist – hierin besteht eine Analogie zu den zuvor betrachteten Modellen – eine geeignete Relativierung auf eine epistemische Situation notwendig. Die Einschränkung der möglichen Ursachen wird durch eine Relativierung auf einen Fragekontext – einen Kontrast – erreicht. Die grundlegende Frage, auf die eine solche Erklärung eine Antwort gibt, lautet: „Warum ist $E$ im vorliegenden Fall $e_1$ eingetreten, aber nicht in den Fällen $e_2$, $e_3$, $e_4$, ...?“ Eine Antwort nennt einen oder mehrere kausal relevante Faktoren, die in dem positiven Fall gegeben und in den negativen Fällen nicht gegeben waren. Dieses Erklärungsmodell ist durch die im Prinzip beliebig wählbare *Kontrastklasse* relativiert auf eine Fragestellung.

Diese Erklärungspraxis folgt einem allgemeinen Prinzip: *Unterschiede in der Wirkung sind durch Unterschiede in den Ursachen zu erklären.* Wenn also ein Fall vorliegt, in dem die Wirkung eingetreten ist und ein Fall, in dem sie nicht eingetreten ist, dann wird mindestens ein kausal relevanter Unterschied zwischen den Fällen verlangt. Umgekehrt bedeutet dies, daß als Erklärung nichts in Betracht kommt, in dem die positiven und negativen Instanzen *übereinstimmen.* Dadurch wird eine enorme Vereinfachung des Erklärungsproblems erzielt, da sowohl zahllose triviale als auch zahllose exotische mögliche Ursachen deshalb aus der Menge der möglichen Erklärungen herausfallen, weil sie keinen relevanten Kontrast bilden, d.h. die positiven und negativen Instanzen sich nicht in diesen Merkmalen unterscheiden.

Der Begriff der Konstrastklasse findet sich bei einer Anzahl von Autoren, darunter Scriven (1964, 1975), Mackie (1980) und insbesondere Van Fraassen (1980). Diese Ansätze unterscheiden sich aber in der Definition der Kontrastklasse und der Bestimmung des Begriffs der Kausalität erheblich von dem vorliegenden Ansatz. Am nähesten verwandt ist der vorliegende Ansatz mit dem Modell Scrivens (1964, 1975). Scriven versucht jedoch, den Begriff der Erklärung auf Grundlage des Ursachenbegriffs zu bestimmen, während ich umgekehrt den Begriff der Erklärung mit Hilfe des Begriff des Verursachung erkläre.

Gegenüber den Eingangs dargestellten Ansätzen hat der vorliegende Ansatz zwei wesentliche Vorzüge:

1. Das Wissen um $E$ spielt eine wesentliche Rolle. Ohne $E$ ließe sich die Elimination nicht durchführen, da nur das vorausgesetzte Wissen um $E$ uns zu rechtfertigen erlaubt, daß eines der $C_i$ gegeben sein muß.

2. Es ist eine Relativierung auf das *tatsächliche* Wissen des Fragenden möglich, wodurch die Problematik der Überzeugungsrevision umgangen wird.

Dadurch ist dieser Ansatz wesentlich natürlicher und einfacher.

Der Ansatz muß auf zwei grundlegende Fragen eine Antwort geben: a) Wie ist der Begriff der Kausalität zu bestimmen? b) wie wird die Relativierung auf einen Fragekontext erzielt? Eine Antwort soll im folgenden Abschnitt skizziert werden. Einzelheiten zu beiden Fragen behandle ich in May (1999), wo auch die übrigen Ansätze genauer diskutiert werden.

## 3.1 Minimale Theorie

Der Aspekt der Regelhaftigkeit unseres kausalen Wissens wird durch das Konzept der minimalen Theorien erfaßt. Zugrunde liegen die Begriffe der hinreichenden und der notwendigen Bedingung in ihrer üblichen logischen Verwendung, sowie ein Konzept der Minimalität, welches benötigt wird, um den Aspekt der kausalen Relevanz modellieren zu können. Das Konzept der minimal hinreichenden Bedingung zur Analyse der kausalen Relevanz wurde von verschiedenen Autoren, insbesondere aber John Mackie (1980) verwendet.

**Minimal hinreichende Bedingung** Eine konjunktive hinreichende Bedingung $A$ von $E$ ist eine minimal hinreichende Bedingung genau dann, wenn kein echter Teil von ihr hinreichend ist. Ein echter Teil von $A$ ist eine Konjunktion $A'$, die weniger Konjunkte als $A$ enthält, wobei jedes Konjunkt in $A'$ ein Konjunkt in $A$ ist.

**Minimal notwendige Bedingung** Eine disjunktive notwendige Bedingung $B$ von $E$ ist eine minimal notwendige Bedingung genau dann, wenn kein echter Teil von ihr notwendig ist. Ein echter Teil von $B$ ist eine Disjunktion $B'$, die weniger Disjunkte als $B$ enthält, wobei jedes Disjunkt in $B'$ ein Disjunkt in $B$ ist.

**Minmale Theorie** Eine minimale Theorie für $E$ ist eine minimal notwendige Bedingung von minimal hinreichenden Bedingungen für $E$.

Ein Kandidat für eine partielle Ursache ist nun ein Faktor, welcher Teil einer minimalen Theorie für das zu erklärende Phänomen ist (in Mackies Terminologie eine INUS-Bedingung). Eine minimale Theorie berücksichtigt, daß es alternative Ursachen für ein gegebenes Phänomen geben kann, und sie berücksichtigt, daß jede Ursache normalerweise ein umfangreiche Menge von positiven und negativen Bedingungen enthält. In dem hier vorgestellten Modell werden minimalen Theorien stets auf einen Kontext relativiert. Sie sind also nicht schlechthin hinreichend oder notwendig. Der Definition dieses Kontextes dienen die folgenden Definitionen.

### 3.2 Epistemische Relativierung

**Kontrastklasse** Die Kontrastklasse $K$ ist eine Menge von Instanzen $K = K^+ \cup K^- = \{e_1, e_2, ..., e_n\}, K^+ \neq \emptyset, K^- \neq \emptyset$.

Eine Kontrastklasse von $E$ ist einfach die Vereinigung zweier beliebiger nicht-leerer Mengen positiver und negativer Instanzen des zu erklärenden Phänomentyps. Durch die Kontrastklasse wird die Fragestellung definiert, indem bestimmte Faktoren aus der Betachtung ausgeschlossen werden oder hereingenommen werden; sie definiert konstante Faktoren, die bei der Suche nach einer Erklärung keine Berücksichtigung finden. Sie sorgt im Zusammenspiel mit dem folgenden Begriff für die gewünschte epistemische Relativierung.

**Kausales Feld** Sei $A_i$ ein Merkmal oder seine Negation. Das kausale Feld ist die Menge derjenigen Merkmale $A_i$, die allen Elementen $e_i$ der Kontrastklasse $K$ zukommen:

$$F_K = \{A_i | (\forall e)(e \in K \rightarrow e \in A_i)\}.$$

Was allen Elementen zukommt, macht keinen Unterschied zwischen den positiven und negativen Instanzen der Wirkung und kann deshalb nicht zur Erklärung herangezogen werden. Dadurch vereinfacht sich das Erklärungsproblem.

**Generalisierte Kontrastklasse** Die Generalisierte Kontrastklasse $K_F^*$ relativ zu einem kausalen Feld $F_K$ ist die Menge aller Objekte, welchen sämtliche Merkmale des Feldes zukommen:

$$K_F^* = \{e_i | (\forall A)(A \in F_K \rightarrow e_i \in A)\}.$$

Mit der Bestimmung der Kontrastklasse und des Feldes über die Extension steht ein Verfahren zur Verfügung, den Bezug auf den Kontext zu begrifflich präzisieren, auch wenn das Feld *praktisch* wohl meist durch solche Wendungen wie unter „normalen Umständen“, „im gesunden Körper“ etc. fixiert wird.

### 3.3 Kausale Relevanz

**Kausale Relevanz** Sei Faktor $C$ Teil einer minimalen Theorie $M$ für $E$ relativ zur Kontrastklasse $K$. Sei $K'$ eine beliebige Kontrastklasse, so daß $K$ eine Teilmenge von $K'$ ist. Dann ist $C$ *kausal relevant* für $E$ genau dann, wenn es für alle $K'$ eine minimale Theorie $M'$ für $E$ gibt, so daß $C$ Teil von $M'$ relativ zu $K'$ ist.

Ein Faktor ist nur dann kausal relevant, wenn er unter Erweiterungen der Kontrastklasse durch zusätzliche Instanzen Teil einer minimalen Theorie für die Wirkung bleibt. *Durch Erweiterung der Kontrastklasse geht kausale Relevanz niemals verloren.* Diese Invarianz und der Mechanismus der Erweiterung der Kontrastklasse erlauben es, verschiedene partielle Kausaltheorien zu einer umfassenderen Theorie zu synthetisieren. Unser kausales Wissen ist also in der Hinsicht monoton, daß sich eine Theorie zwar immer weiter ausdifferenzieren läßt, indem die Kontrastklasse erweitert wird, die einmal als kausal relevant identifizierten Faktoren dabei aber immer noch vorkommen. Stellt sich heraus, daß ein Faktor nicht mehr vorkommt, dann war er kein kausal relevanter Faktor.

## 4 Kontextabhängigkeit

Nun soll gezeigt werden, wie sich die eingeführten Begriffe nutzen lassen, um das Problem der Kontextabhängigkeit von Erklärungen in den Begriff zu bekommen.

Wenn, um ein in ähnlicher Form oft angeführtes Beispiel zu nehmen, nach der Ursache eines Autounfalls gefragt wird, dann wird der Kfz-

Sachverständige vielleicht auf den schadhaften Zustand der Bremse verweisen; der Mediziner wird die Beeinträchtigung des Fahrvermögens durch Alkoholgenuß anführen; und ein dritter Gutachter wird die verminderte Bodenhaftung der Reifen infolge von Nässe als Ursache herausstellen.

Was der jeweilige Gutachter als *die Ursache* herausgreift, ist abhängig von seiner Fragestellung und dem jeweiligen Kontext. Damit scheint es relativ willkürlich zu sein, ob etwas als Ursache gilt oder nicht, da nur die Fragestellung und der Kontext entsprechend gewählt werden müssen. Ich möchte nun zeigen, wie sich das Phänomen der Kontextabhängigkeit im vorliegenden Ansatz in einer Weise lösen läßt, die eine gemeinsame Basis für die verschieden Erklärungen erkennbar werden läßt.

Die verschiedenen Gutachter antworten auf unterschiedliche Kausalfragen und unterstellen verschiedene Kontrastklassen. Der Kfz-Sachverständige relativiert sein Gutachten implizit auf Fälle, in denen die gleichen Straßenverhältnisse vorlagen, so daß der Zustand der Bremsen der relevante Unterschied ist. Der Mediziner relativiert auf Fälle, in denen es bei vergleichbarem Wagenzustand und ähnlichen Straßenverhältnissen zu keinem Unfall kam; und der dritte Gutachter bezieht sich auf Fälle, in denen die Straße trocken war, so daß trotz Alkoholgenuß und schadhaften Bremsen kein Unfall geschah.

Indem jeder der Gutachter eine solche Relativierung vornimmt, werden die Faktoren der jeweils anderen Gutachter Teil des kausalen Feldes, da sie in der jeweiligen Kontrastklasse konstant sind. Deshalb können sie von dem jeweiligen Gutachter ignoriert werden und er kann sich auf den Aspekt des Unfalls konzentrieren, für den er zuständig ist.

Die Lösung basiert auf der Forderung der Invarianz der kausalen Relevanz. Die Definition der kausalen Relevanz ist wesentlich durch den Mechanismus des Wechselns der Kontrastklasse, der damit verbundenen Erzeugung unterschiedlicher kausaler Felder und der Notwendigkeit, in diesem Wechsel eine Konstanz oder Invarianz zu finden, motiviert.

Jeder der Gutachter macht eine Kausalbehauptung, die nur dann zutreffend ist, wenn der behauptete Faktor unter Erweiterung der Kontrastklasse relevant bleibt. Da eine solche Erweiterung beliebig geschehen darf, kann insbesondere auch die Vereinigung $K = K_1 \cup K_2 \cup K_3$ der jeweils unterstellten Kontrastklassen genommen werden. Relativ zu dem kausalen Feld, welches durch die Kontrastklasse $K$ definiert wird, muß jeder der von den Gutachtern genannten Faktoren Teil einer minimalen Theorie für $E$ sein. Dadurch ergibt sich sozusagen das Gesamtgutachten, in dem die unterschiedlichen Einzelaussagen integriert sind.

Stellt sich heraus, daß ein Faktor relativ zu $K$ nicht Teil einer minimalen Theorie für $E$ ist, dann ist dieser Faktor nicht kausal relevant. Dies kann etwa der Fall sein, wenn ein Gutachter in seinem Feld eine gemeinsame Ursache $C$ zweier Faktoren $E_1$ und $E_2$ konstant gehalten hat und fälschlicherweise eine Kausalbeziehung zwischen $E_1$ und $E_2$ konstruiert hat. Sobald durch Erweiterung der Kontrastklasse $C$ variabel wird, erweist sich diese Beziehung als redundant.

# 5 Ausblick: Erklärung und Experiment

Wie läßt sich der dargestellte Ansatz konstruktiv für die Suche nach Erklärungen nutzen? Die Ermittlung einer minimalen Theorie relativ zu einer Kontrastklasse ist nicht trivial. Ich beschreibe die Suche nach Ursachen im experimentellen Kontext mittels der Technik der eliminativen Induktion ausführlich in Graßhoff & May (1995), May (1998, 1999). Es besteht ein enger Zusammenhang zwischen einem Experiment und der kausalen Erklärung: Die Ermittlung von Ursachen im Experiment ist die Suche nach Erklärungen, und die Kontrolle und systematische Variation von Faktoren dient der partiellen Fixierung des kausalen Feldes. Die verschiedenen konkreten Ergebnisse eines Versuches entsprechen den Elementen der Kontrastklasse, durch die sich das gesamte Feld konstituiert. In einer Anwendung der dargestellten Theorie auf das *Maschinelle Entdecken* zeige ich in den genannten Arbeiten, wie sich durch eine geschickte Wahl des Feldes in einem iterativen Verfahren inkrementell eine komplexe Kausaltheorie bilden läßt. Dies wiederum entspricht dem Wechsel der Kontrastklasse, wie sie bei den Gutachtern beschrieben wurde.

Der Fragende ist frei, eine Kontrastklasse zu wählen; unterschiedliche Kontrastklassen halten unterschiedliche Faktoren konstant und lassen unterschiedliche Faktoren als Antwort zu. Eine gu-

te Frage zu stellen, heißt häufig, eine interessante Kontrastklasse zu wählen. Eine gute Frage impliziert oft einen scharfen, relevanten Kontrast. Indem eine gute Frage einen scharfen Kontrast wählt, werden viele irrelevante Faktoren ausgeschieden und eine Antwort läßt sich leichter finden. Eine unpräzise Frage hingegen wählt einen falschen Kontrast, der zuviele Antworten zuläßt. Auch hier besteht ein enger Zusammenhang zwischen der Kunst, eine gute Frage zu stellen und der Kunst, ein gutes Experiment zu planen.

## Literatur


Boutilier, C. und V. Becher (1995). Abduction as belief revision. *Artificial Intelligence 77*, 43–94.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Chajewska, U. und J. Halpern (1997). Defining explanation in probabilistic systems. In: *Uncertainty in Artificial Intelligence*, Volume 13, S. 62–71.

Gärdenfors, P. (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Cambridge (Mass.): MIT Press.

Geffner, H. (1992). *Default Reasoning. Causal and Conditional Theories*. Cambridge (Mass.): MIT Press.

Graßhoff, G. und M. May (1995). From historical case studies to systematic methods of discovery. In: R. Valdes-Perez (Hrsg.), *AAAI Spring Symposium Series "Systematic Methods of Discovery", Stanford University*, Menlo Park. AAAI-Press.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. New York: The Free Press.

Hempel, C. G. und P. Oppenheim (1965[1948]). Studies in the logic of explanation. In: C. G. Hempel (Hrsg.), *Aspects of Scientific Explanation*, S. 245–290. New York: The Free Press.

Howson, C. und P. Urbach (1993). *Scientific Reasoning. The Bayesian Approach*. 2. erw. Aufl. Chicago: Open Court.

Mackie, J. L. (1980[1974]). *The Cement of the Universe. A Study of Causation*. 2. Aufl., korrigiert und mit neuem Vorwort, 1. Aufl 1974. Oxford: Clarendon Press.

May, M. (1998). Causation and discovery. In: N. Gilbert und P. Ahrweiler (Hrsg.), *Computer Simulation in Science and Technology Studies*. Berlin: Spinger.

May, M. (1999). *Eine Theorie des kausalen Schließens*. Berichte des Graduiertenkolleg Kognitionswissenschaft Universität Hamburg Nr. 64. Hamburg: Universität Hamburg. (gleichzeitig Dissertation).

Mitchell, T. (1996). *Machine Learning*. New York: McGraw-Hill.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo: Morgan Kaufman.

Pearl, J. (1999). *Simpson's Paradox: An Anatomy*. UCLA. Technischer Bericht.

Reichenbach, H. (1949). *Theory of Probability*. Berkeley: University of California Press.

Salmon, W. (1971). *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: University of Princeton Press.

Scriven, M. (1964). The structure of science. *Review of Metaphysics 17*, 403–424.

Scriven, M. (1975). Causation as explanation. *Noûs 9*, 3–15.

Srinivasan, A. (1999). A study of two sampling methods for analyzing large data sets with ILP. *Data Mining and Knowledge Discovery 3(1)*, 95–123.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.