PlexPlain

Erklärende KI für Komplexe Lineare Programme

am Beispiel intelligenter Energiesysteme

TECHNISCHE UNIVERSITÄT DARMSTADT

Centre for Cognitive Science

Abschließender Sachbericht

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS19081 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor*innen.	
Zuwendungsempfänger:	Technische Universität Darmstadt
Laufzeit des Vorhabens:	01.04.2020 – 31.07.2023
Autoren des Berichts:	Jonas Hülsmann, Inga Ibs, Claire Ott, Matej Zecevic, Dirk Balfanz, Frank Jäkel, Kristian Kersting, Constantin Rothkopf, Florian Steinke

Version 8 – 14.12.2023

Inhalt

1 Kurzdarstellung

1.1 Aufgabenstellung

Technische und soziale Systeme werden immer komplexer. Künstliche Intelligenz (KI) hat das Potenzial, Planungen und Entscheidungen in solchen komplexen Systemen essentiell zu unterstützen. Durch KI und maschinelles Lernen gewonnene Vorhersagen und Handlungsstrategien für komplexe Systeme entziehen sich bisher einfachen und transparenten Erklärungen. In vielen Anwendungssituationen ist es allerdings unverzichtbar, gut erfassbare aber dennoch sachlich begründbare Erklärungen für die Ergebnisse der KI-Systeme geben zu können, z. B. um politische Entscheidungen zu rechtfertigen, Bürgerbeteiligung zu ermöglichen oder Therapieempfehlungen verständlich zu machen.

Hintergrund von „PlexPlain“ ist, dass lineare Programmierung, also die Optimierung von Zielfunktionen auf Basis linearer Gleichungen, eine grundlegende KI-Methode ist, die sehr häufig für Optimierung und Planung in komplexen Systemen eingesetzt wird. Auch aktuelle Vorhersagemethoden, die auf neuronalen Netzen beruhen und das sogenannte Verstärkungslernen lassen sich durch lineare Programme analysieren. In Anwendungen können diese Millionen von Variablen enthalten, deren Zusammenwirken oft auch für Experten nur schwer verständlich ist. Damit bilden lineare Programme eine große Klasse an Problemen, für die dringend kognitiv adäquate Erklärungen benötigt werden.

Ziel für „PlexPlain“ war daher zum einen grundlagenorientiert in Verhaltensstudien zu untersuchen, wie Menschen lineare Programme verstehen. Zum anderen sollten neue Methoden entwickelt werden, die lineare Programme (semi-)automatisch vereinfachen, in graphische Modelle übersetzen und Erklärungen erzeugen. In Verhaltensstudien sollten auch Versuchspersonen mit Planungsproblemen konfrontiert werden, die durch lineare Programme optimal gelöst werden können. Die zu beobachteten Lösungsstrategien sollten in die Entwicklung der neuen Methoden einfließen.

Das Projekt war eine interdisziplinäre Zusammenarbeit zwischen Kognitionswissenschaft, KI sowie der Elektrotechnik. Als Anwendungsfälle dienen dabei u. a. komplexe Modelle von Energiesystemen im Rahmen der Planung der Energiewende (siehe z. B. Abb. 1), die in Abstimmung mit den assoziierten Partnern Siemens AG und Entega AG untersucht wurden.

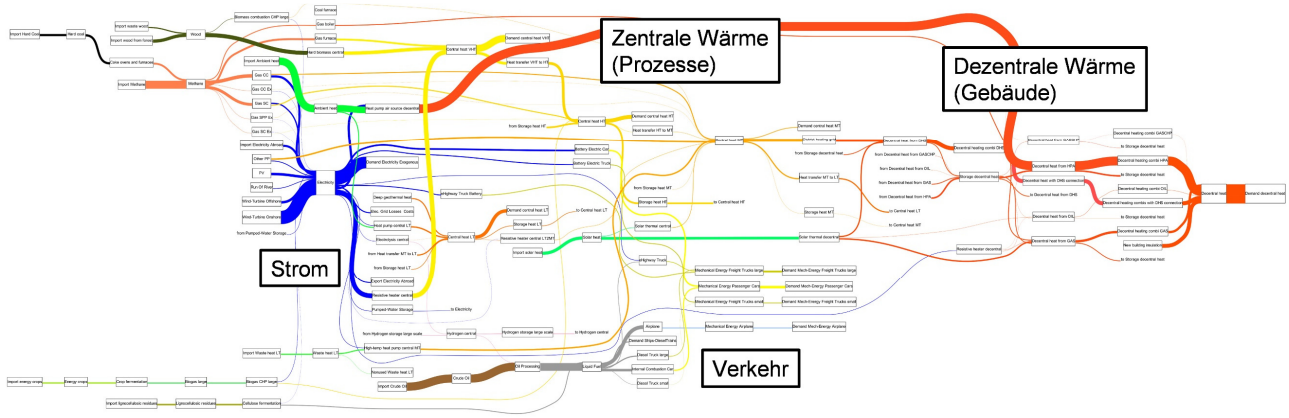


Abbildung 1: Sankey-Diagramm / Energieflüsse zwischen verschiedenen Energieformen für ein multimodales Deutschlandmodell (Strom, Wärme, Verkehr) in 2050

1.2 Anknüpfung an den wissenschaftlichen und technischen Stand

Zu den wesentlichen Grundannahmen der Explainable AI (erklärbaren KI) gehört, dass transparentere, besser durch den Menschen interpretierbare KI-Systeme das Verständnis solcher Systeme verbessern und das Vertrauen in sie stärken. Die informierte Nutzung von KI-Systemen hat außerdem einen positiven Einfluss auf das Gesamtergebnis. Statt den Menschen zu ersetzen gilt das Motto: Der Mensch nutzt KI. Diese Zusammenhänge sind aus unterschiedlichen Perspektiven gut untersucht und belegt. Zwei Ansätze sind dabei generell zu unterscheiden: (1) Das Transparentmachen von Ursachen und Abhängigkeitsketten, welche die Ergebnisse der KI bestimmen. Hier geht es um die Erhöhung der Interpretierbarkeit (Interpretability) der Wirkungsweise der KI durch Offenlegen der kausalen Mechanismen. (2) Die Erklärung (Explanation) von konkreten KI-Ergebnissen durch Angabe von Gründen. Diese Gründe können aus Modellen wie unter (1) abgeleitet sein, aber auch durch andere

Mechanismen (z. B. mittels Sensitivitätsanalysen) erzeugt werden. Während im ersten Fall ein Fokus auf Fragen der korrekten Modellierung und Vereinfachung komplexer KI-Systeme vorherrscht, ist im zweiten Ansatz eine Betonung kognitionswissenschaftlicher Fragen zu finden, d.h. die Betrachtung, was gute Erklärungen für Menschen ausmacht. Beide Richtungen wurden in „PlexPlain“ verfolgt.

1.3 Ablauf des Vorhabens

Die Projektlaufzeit von PlexPlain betrug 40 Monate, inkl. einer 4-monatigen kostenneutralen Verlängerung. Die Projektarbeit war parallel organisiert in 4 inhaltlichen Schwerpunkten bzw. Arbeitspaketen (AP): AP 1 „Abstraktion erklärender Variablen“, AP 2 „Extraktion von Wirkzusammenhängen“, AP 3 „Konstruktion zielbezogener Begründungen“ und AP 4 „Anwendung und Plausibilisierung im Energiesektor“. Entlang der Meilensteine (MS) des Projektes wurden zunächst Grundlagen untersucht und durch erste Versuche Hypothesen gebildet (MS 1, 28.10.2020), die Ansätze zu diesen Schwerpunkten schrittweise aufgebaut und entwickelt (MS 2, 02.11.2021 und MS 3, 31.03.2023), sowie schließlich anhand von Beispielen sowie z. T. mit Proband:innen getestet und mit den Industriepartnern diskutiert und bewertet (MS 4, 17.07.2023).

1.4 Wesentliche Ergebnisse und Forschungszusammenarbeit

Ziel des Projektes war die Entwicklung von (semi-)automatischen Methoden zur Erstellung von kognitiv adäquaten Erklärungen der Ergebnisse von linearen Programmen und deren Anwendung auf Energiesysteme. Eine grundlegende Frage war deshalb, wie kognitiv adäquate Erklärungen für lineare Programme aussehen können. Es wurden verschiedene Szenarien (in der Form von Computerspielen) entwickelt, in denen Versuchspersonen lineare Programme lösen mussten und dabei erklärten, wie sie vorgingen und warum ihre Lösung gut ist. Dabei zeigte sich, dass kognitiv adäquate Erklärung nicht-zyklisch sind und schrittweise vorgehen. Um zyklische Abhängigkeiten zu vermeiden, nutzten Versuchspersonen verschiedene Heuristiken, Variablen zu ordnen und dann deren Einfluss und Werte nacheinander zu erklären. Ein wesentliches empirisches Ergebnis ist eine Beschreibung der menschlichen Heuristiken. Diese und die in den Experimenten beobachteten Lösungsstrategien wurden dann genutzt, um zwei Ansätze für post-hoc Erklärungen von Lösungen zu entwickeln. Der erste führte zum Werkzeug SimplifEx. Dies beschreibt Lösungen und beruht auf iterativen lokalen Erklärungen sowie klassischen Vorverarbeitungsmethoden zur Vereinfachung von linearen Programmen. Der zweite beruht auf einer post-hoc Zusammenfassung der Lösungsstrategie durch ein logisches Programm, das aus kognitiv plausiblen Bausteinen besteht, die aus den Erklärungen der Versuchspersonen gewonnen wurde. Beide Ansätze führten zu Computerprogrammen, die auf den von uns getesteten Beispielen gut funktionieren, und nun in anderen Anwendungsdomänen erprobt werden können. So haben wir z. B. SimplifEx auch schon auf einem klassischen Problem der Ernährungsplanung getestet. Ein offenes Problem bei der Anwendung auf Energiesysteme ist noch der Effekt von zeitlichen Abhängigkeiten, der zu sehr komplexen Erklärungen führt. Ein weiteres Ergebnis ist, dass die Extraktion von Wirkzusammenhängen aus linearen Programmen mit klassischen Methoden des kausalen Lernens, nicht gut funktioniert, weil die Lösung in linearen Programmen potentiell zyklische Abhängigkeiten hat und diese insbesondere bei Energiesystemen eine wichtige Rolle spielen. Zu einem gewissen Grad ist es uns trotzdem gelungen, neue Methoden zu entwickeln, die kausale Beziehungen aus linearen Programmen extrahieren. Obwohl diese Methoden noch nicht für alle Aspekte von Energiesysteme passend sind, weil auch hier Zeitreihen noch nicht adäquat verarbeitet werden können, konnten wir ihre Funktionsweise in anderen Anwendungsdomänen demonstrieren. Speziell für Energiesysteme haben wir außerdem klassische Sensitivitätsanalysen, wie sie in der Praxis bereits eingesetzt werden, erweitert. Dafür haben wir Methoden aus der Explainable AI (wie z.B. LIME) zur Analyse von linearen Programmen angepasst und konnten zeigen, dass man so komplexere Energiesysteme intuitiver als mit vorher existierenden Methoden analysieren kann.

Da alle beteiligten Wissenschaftler:innen an der TU Darmstadt verortet waren, konnte das Team trotz der Einschränkungen durch die COVID-19 Pandemie durchgängig eng integriert arbeiten. Die interdisziplinären Verbindungen zwischen Kognitionswissenschaft, Informatik und Elektrotechnik wurden durch die Zusammenarbeit über drei Fachbereiche durch das Projekt gestärkt und legen so die Grundlage für eine andauernde, vertiefte interdisziplinäre Forschungsarbeit. Die assoziierten Partner Siemens und Entega haben das Projekt begleitet und beraten und haben die hohe Relevanz der Forschungsfragen des Projektes für die Anwendung bestätigt.

2 Eingehende Darstellung

Die wesentlichen Arbeitshypothesen in PlexPlain waren, dass (1) kognitiv adäquate Erklärungen von der großen Zahl an Variablen abstrahieren müssen, (2) abstrakte Wirkzusammenhänge für das menschliche Verständnis vorteilhaft in einem graphischen Modell dargestellt werden können und (3) Erklärungen zielbezogen sein müssen, d. h. es muss begründet werden, warum eine Entscheidung im Hinblick auf das zu erreichende Ziel gut und besser als eine andere ist.

Entsprechend widmete sich AP 1 der „Abstraktion erklärender Variablen“ und der Fragestellung, wie Menschen zu für sie nachvollziehbaren Abstraktionen kommen und wie ggfs. solch ein Vorgang formalisierbar ist. AP 2 „Extraktion von Wirkzusammenhängen“ untersuchte, ausgehend von bereits komplexitätsreduzierten Systemen, wie Wirkmechanismen aus gegebenen Modellen extrahiert und graphisch dargestellt werden können. Es wurden neue KI-Methoden erforscht, die (semi-)automatisiert graphische Modelle und zielbezogene Erklärungen aus komplexen linearen Programmen erstellen. Wir haben zudem untersucht wie für Menschen kognitiv adäquate Erklärungen aussehen und wie diese erreicht werden können. Da komplexe lineare Programme in der Regel Zielfunktionen optimieren, kann sich eine Erklärung dieser jedoch nicht auf Wirkzusammenhänge beschränken, sondern muss Begründungen anbieten, warum die optimalen Handlungen zur Erreichung des Ziels besser als andere plausible Handlungen sind. Dies erforschte das AP 3 „Konstruktion zielbezogener Begründungen“. Im AP 4 ging es um eine Anwendung der untersuchten Methoden in einer konkreten Domäne, es erfolgte die „Anwendung und Plausibilisierung im Energiesektor“. Ebenfalls betrachtet wurde die mögliche Übertragbarkeit unserer Lösungen auf andere Anwendungsbereiche. AP 5 war schließlich organisatorisch der „Projektkoordination und Abstimmung“ gewidmet.

Nachfolgend werden die wichtigsten Ergebnisse aus diesen Arbeitspaketen genauer besprochen. Für eine detaillierte Darstellung auf wissenschaftlicher Ebene wird auf die Fachpublikationen des Projektes verwiesen (siehe Abschnitt 2.6).

2.1 Verwendung der Zuwendung und erzielte Ergebnisse

AP 1	Abstraktion erklärender Variablen
	Es wurden verschiedene Experimente durchgeführt, um menschliches Verhalten beim Lösen linearer Programme (LP) zu analysieren (siehe auch AP 3). Bei der hier vorgestellten Studie ging es um die Frage, welche und wie viele verschiedene Informationen für verschiedenen Lösungsstrategien verwendet werden. Außerdem wurde betrachtet, wie verschiedene Lösungen verglichen und bewertet werden. Die Experimente beruhen auf einem einfachen LP, der Furniture Factory, bei der es darum geht mit gegebenen Rohstoffen Möbelstücke zu bauen und so den erzielbaren Profit zu maximieren. Dieses Szenario wurde von uns entwickelt, da es kein domänenspezifisches Vorwissen benötigt oder durch solches konfundiert werden könnte und die Komplexität angemessen ist. Aus diesem Szenario wurden im Rahmen einer Bachelorarbeit [A1] zwei Spiele konzipiert. Eines dieser Spiele erlaubt die Exploration des gesamten Lösungsraums und wurde für ein Laborexperiment verwendet, bei dem Testpersonen durchgängig alle ihre jeweils ausgeführten Aktionen erklären und begründen sollten (siehe Abbildung 1). Die Studie bietet tiefere Einblicke in das Verständnis und das interne Modell der Proband:innen im Umgang mit linearen Programmen, besonders hinsichtlich der Exploration des Lösungsraums. Die Ergebnisse der Studie zeigen, dass viele Proband:innen das Problem reduzieren indem sie entweder primär auf das Optimierungskriterium oder die Beschränkungen achten [p1]. Außerdem gibt es unterschiedliche Strategien, die verschiedenen Lösungen zu vergleichen und zu bewerten [p3]. Die Struktur der Furniture Factory wurde mit der anderer Problemlöseaufgaben verglichen und in den größeren Kontext der vorhandenen Literatur gesetzt, um

Ergebnisse besser transferieren zu können und allgemeinere Aussagen über LPs in der Problem-Lösungs-Forschung zu treffen.

Um lineare Programme nachvollziehbar zu vereinfachen, wurden vorhandene Verfahren analysiert, welche bis jetzt primär zur Beschleunigung von Lösungsverfahren eingesetzt wurden. Einige dieser Methoden wurden um Feedback erweitert, welches die Vereinfachungen erklärt und somit wichtige Informationen über das ursprüngliche und verbleibende LP gibt.

Abbildung 2: Beispielstatus des Experiments Furniture Factory

SimplifEx [p2] ist ein Tool zur automatischen Generierung kognitiv adäquater Erklärungen für die optimale Lösung eines LPs. Das Tool besteht aus drei Grundbausteinen, welche in mehreren Iterationen auf das LP angewendet werden (siehe Abbildung 3).

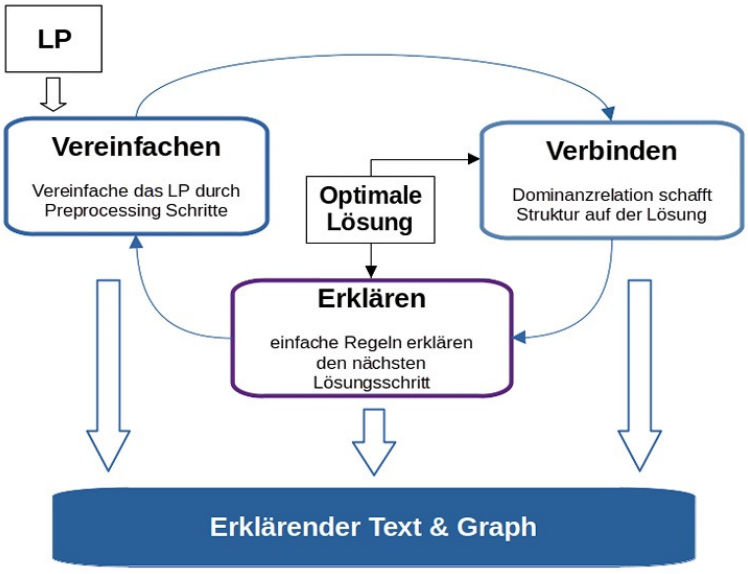


Abbildung 3: Schematische Darstellung des Tools SimplifEx zur Vereinfachung und Erklärung bestimmter LPs

Die schriftliche Detaildarstellung wird durch einen Graph erweitert, welcher sowohl die Dominanzrelation in der optimalen Lösung zeigt, als auch die Heuristiken der einzelnen Schritte farblich markiert. Die Entwicklung der Optimierungsfunktion und die sukzessive Erfüllung der Beschränkungen sind ebenfalls in der graphischen Darstellung enthalten.

Zuerst werden bekannte vorbereitende (Preprocessing-)Verfahren verwendet um LPs nachvollziehbar zu vereinfachen. Anschließend werden Variablen mit Hilfe der Dominanzrelation verglichen und nach ihren Werten, d. h. ihrem Lösungsbeitrag im LP geordnet. Auf diese Weise werden das LP und seine optimale Lösung strukturiert. Um diese Methode möglichst gut auszunutzen wurde sie verallgemeinert, um mehr Fälle von Dominanz zwischen Variablen zu umfassen. Im dritten Schritt werden Heuristiken verwendet um für einzelne Variablen der optimalen Lösung Erklärungen zu generieren. Die verwendeten Heuristiken beruhen auf Erkenntnissen der Experimente in AP 1 und AP 3.

Das resultierende Tool wurde unter an verschiedenen Beispielen wie der Furniture Factory und dem Ernährungsproblem von Stigler[1] getestet (Abbildung 4). Außerdem wurde es auf kleine Energiebeispiele angewendet, bei denen eine Merit-Order der Erzeuger herausgearbeitet wurde oder die Veränderung bestimmter Parameter zu anderen Einzelschrittentscheidungen führten (Abbildung 5).

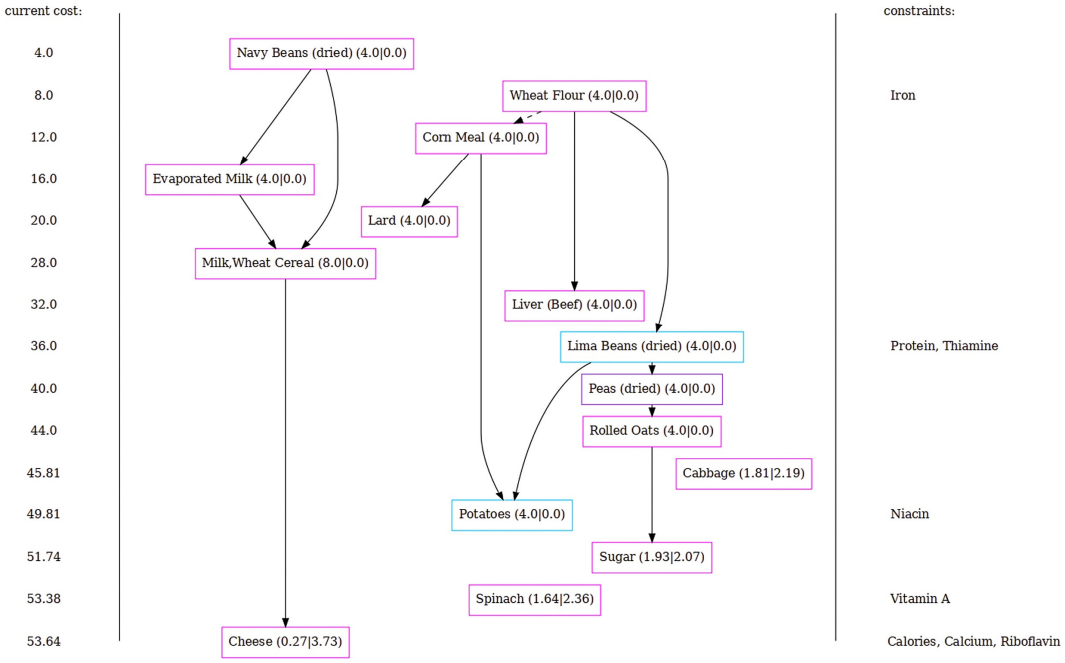


Abbildung 4: Lösungsgraph für Stiglers Ernährungsproblem mit vier als maximal Wert für jedes Nahrungsmittel. Zuerst werden Nahrungsmittel mit sehr guten Nährwerten auf ihren maximalen Wert gesetzt, dann werden übrige Nährstoffbedürfnisse erfüllt.

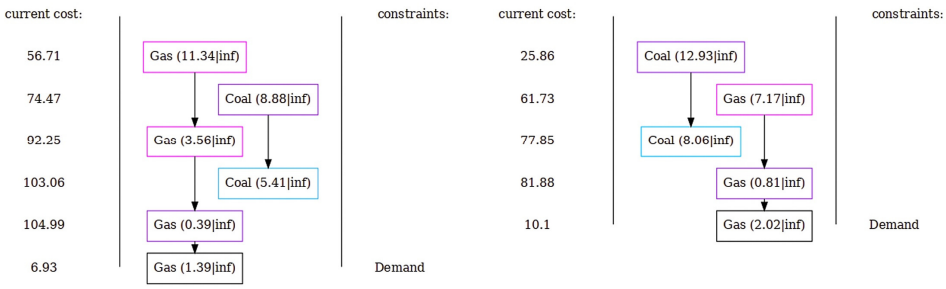


Abbildung 5: Lösungsgraphen für zwei Szenarien eines simplen Energiesystems mit Gas und Kohle. Das Modell soll die Kosten zur Deckung einer fixen Last minimieren und ein CO2-Budget einhalten. In Szenario 1 (Kohle ist günstiger, emittiert aber mehr CO2) wird Gas als erstes gewählt, in Szenario 2 (Gaskraftwerk emittiert weniger CO2 als in Szenario 1) wird Kohle gewählt, da das CO2-Budget einfacher einzuhalten ist.

[1] Stigler, G. J. (1945). The Cost of Subsistence. Journal of Farm Economics

AP 2	Extraktion von Wirkzusammenhängen

Eine wichtige Grundlage der Arbeiten in AP 2 waren Projektvorarbeiten, die zeigen konnten, dass eine Abbildung von LP Optimierungsproblemen in neuronale Netzarchitekturen konsistent möglich ist und Untersuchungen zur Kausalität der Extraktion von Wirkzusammenhängen. Darauf aufbauend, haben wir interventionelle Sum-Product Netzwerke (iSPN) [2] entwickelt. Hierbei werden neuronale Netze in SPN übersetzt zur Abbildung der Konsequenzen von Wirkzusammenhängen (Abbildung 6).

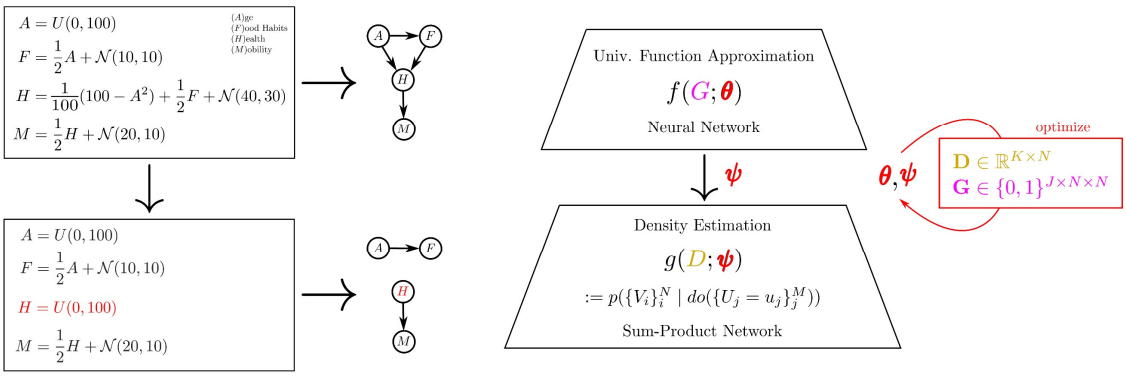


Abbildung 6: Beispiel für die neu entwickelten interventionellen Sum-Product Netzwerke [2]. Zur Analyse von Konsequenzen der Wirkzusammenhänge bildet diese skalierbare Modellklasse neuronale Netzarchitekturen und graphische Modelle auf SPN ab. Dargestellt ist wie medizinische Variablen (Alter (A), Ernährung (F), Gesundheitsbefinden (H) und Mobilität (M)) und explizite Veränderungen dieser (bspw. durch verschriebene Medikation) vom Modell abgebildet werden.

Wiederum spezifisch für LP Optimierungsprobleme haben wir eine mathematische Formalisierung entwickelt, welche quantifizierende Aussagen dazu erlaubt, wie stark eine Änderung eines LP zu einer Veränderung der optimalen Lösung des LPs führt [4]. Die Erklärbarkeit von LP Optimierungsproblemen hängt von den Wirkzusammenhängen ab, die dem LP implizit sind. Daher ist es wichtig die Empfindlichkeit (Sensitivität) der LP-Wirkzusammenhänge gegenüber Veränderungen quantitativ abschätzen zu können (Abbildung 7).

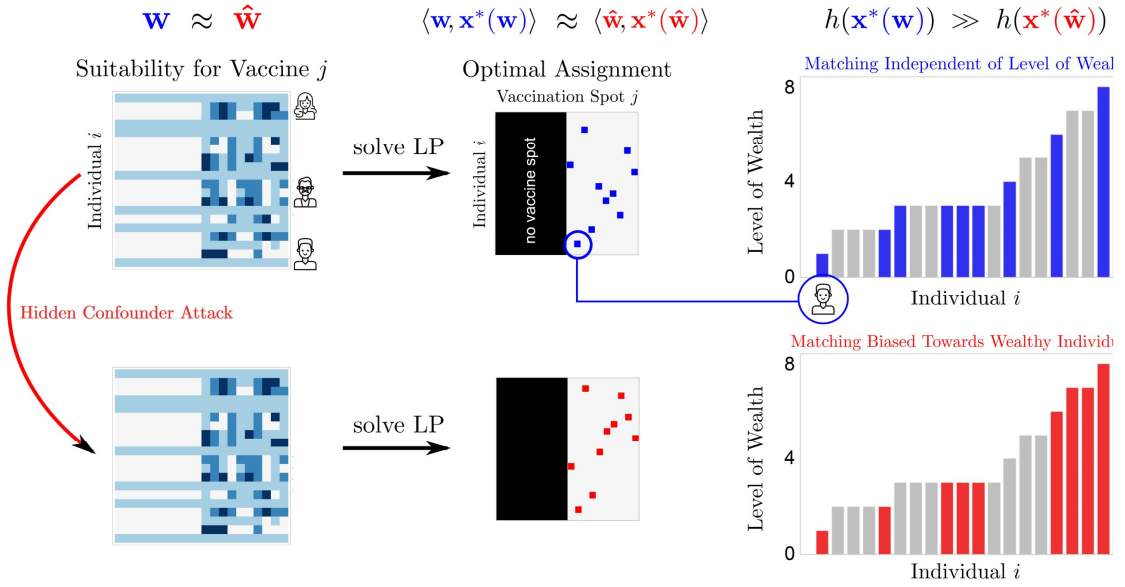


Abbildung 7: Die Übersicht für die Ergebnisse aus [2]. Die Stabilität der Wirkzusammenhänge in LP Optimierungsproblemen wird durch unsere Formalisierung effektiv dargestellt.

Die Ergebnisse von [4] wurden durch die Untersuchung von LPs bzgl. deren Erklärbarkeit durch eine Reihe von Arbeiten erweitert [p6, A4 und p5, A5]. Um den Zusammenhang

zwischen der kausalen Modellierung von LPs und ihrer Erklärbarkeit zu untersuchen, haben wir eine empirische Studie durchgeführt [11]. In dieser Studie wurden Strukturerkennungsmodelle für kausale Wirkzusammenhänge auf Daten angewandt, die ihren Ursprung in einem LP haben (Abbildung 8).

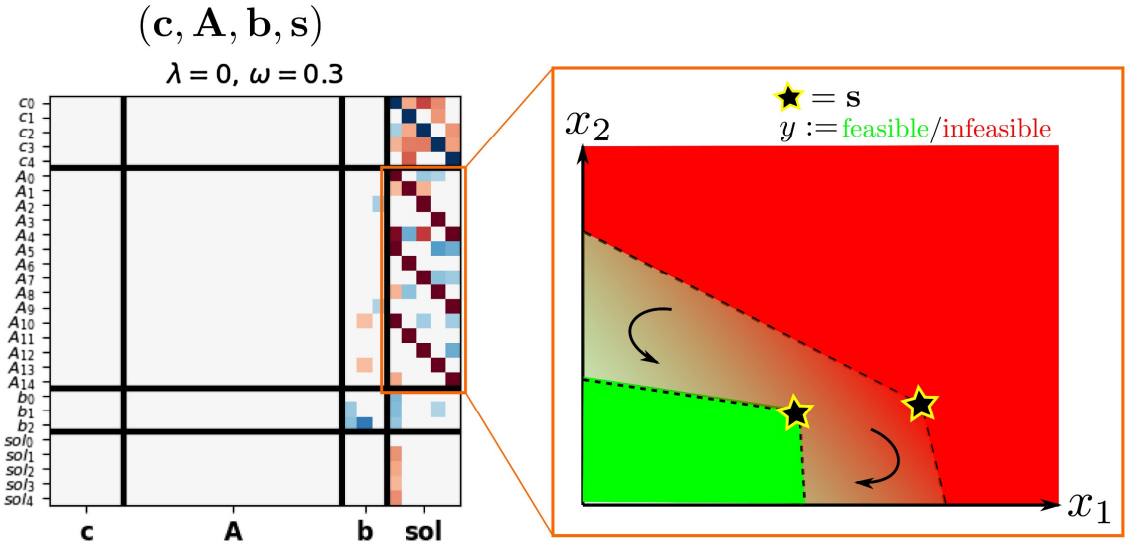


Abbildung 8: Ergebnisse der Studie [11] zeigen, dass die kausalen Wirkzusammenhänge (ohne Berücksichtigung potenziell zyklischer Strukturen), die aus den Daten des LP Daten abgeleitet wurden, konsistent sind zum ursprünglichen LP.

Um automatisiert Erklärungen in natürlicher Sprache zu erstellen, die dem Menschen zunächst bereits per Konstruktion leichter zugänglich sind, haben wir kausale Modelle (wie bspw. das iSPN aus [2]) analysiert und eine Methode zur Erstellung entsprechender Erklärungen entwickelt (genannt Structural Causal Explanations = SCE [P1]). Diese Methode bietet einen potenziellen Grundbaustein für (kausale) Erklärungen von LPs, die zunächst auf kausale Strukturen untersucht wurden wie in der vorherigen Arbeit zu [11]. Des Weiteren zeigt der Vergleich zu existierenden Methoden bedeutende Vorteile bzgl. der Interpretierbarkeit der generierten Antworten (Abbildung 9).

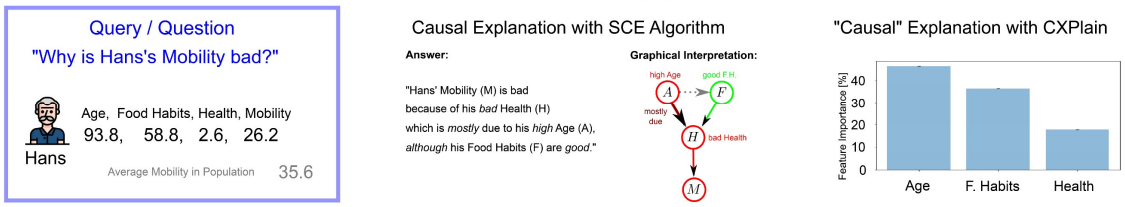


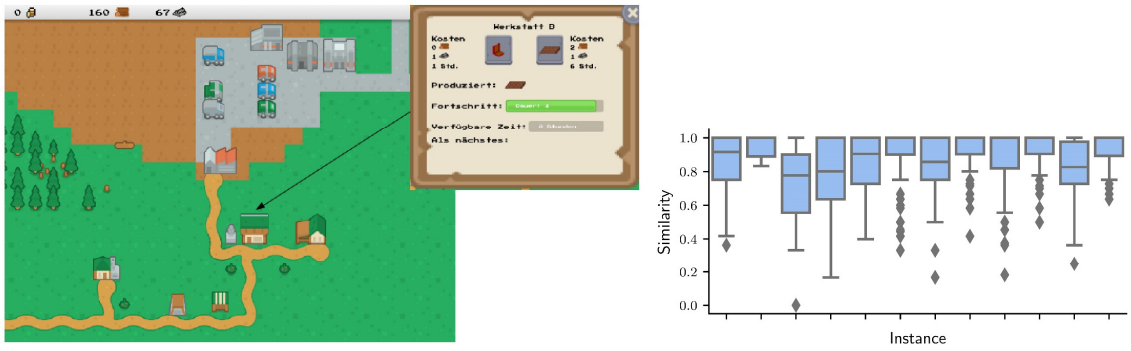
Abbildung 9: Die Übersicht für die Ergebnisse aus [p4], die zeigt dass im Vergleich zu existierenden Methoden zur Erklärung kausaler Wirkzusammenhänge in den zu untersuchenden Daten, der SCE Algorithmus nützlichere Informationen liefert.

Die Ergebnisse der bisher durchgeführten Studie von [p4] wurden anschließend weiter ausgebaut. In [A9] wurden Zeitreihen modelliert um eine direkte Anwendung auf LPs, wie sie bei Energiesystemen zu finden sind, zu ermöglichen. Mit [p7] wurden bislang erste Schritte in Richtung eines besseren theoretischen Verständnisses von neuronal-kausalen Modellen gemacht. Die Erkenntnisse in [p7] können in Zukunft genutzt werden um Ergebnisse aus [2] verbessert auf LPs anzuwenden.

AP 3	**Konstruktion zielbezogener Begründungen**
	Um kognitiv adäquate zielgerichtete Erklärungen für Lösungen von linearen Programmen zu generieren, ist es nicht ausreichend nur kausale Zusammenhänge des zugrunde liegenden Problems einzubeziehen, der Effekt dieser Zusammenhänge auf eine Lösung

muss auch in Form von menschlicher Repräsentation erklärt werden. Dazu gehört zum einen die Problemdefinition in für Menschen verständliche Repräsentationen zu übertragen (AP 1), zum anderen aber auch die einzelnen Teile der Lösung in der bedeutsamen Reihenfolge zu erklären. Hierzu ist es wesentlich zu verstehen wie Problem-definition, Erklärungen und Optimierungsstrategien von Menschen zusammenfinden.

Um dies zu untersuchen wurde das lineare Programm der Furniture Factory (siehe AP 1) im Rahmen einer Bachelorarbeit [A1] in ein neues Experimentalspiel übertragen, in welchem Optimierungsstrategien und post hoc Erklärungen von Menschen untersucht werden können (siehe Abbildung 10 a). In diesem Experiment sollten Proband:innen nacheinander entscheiden welche Möbelstücke sie bauen und planen wie sie mit ihren Aktionen zur optimalen Lösung gelangen. Im Gegensatz zu dem ersten Experimentalspiel der Furniture Factory (siehe AP 1) ging es hierbei nicht darum, den Lösungsraum zu erkunden sondern ohne Rücknahme der Entscheidungen zur optimalen Lösung zu gelangen. Mithilfe dieses Experimentalspiels wurden zwei Studien durchgeführt [p1]. In einer ersten Studie wurden Klickdaten von 31 Versuchspersonen aufgenommen. Desweiteren wurden die Teilnehmer:innen dazu aufgefordert ihre Lösungsstrategien nach Versuchsblöcken (post hoc) zu erklären. Aufgrund dieser Daten wurden Bausteine von Erklärungen und Strategien identifiziert, wie z. B. „wähle das, was am wenigsten Holz verbraucht“, und in eine generellere mathematische Form übertragen [p1]. Desweiteren konnten Ähnlichkeiten zu den Strategien aus dem anderen Experimentalspiel in der Think Aloud Studie festgestellt werden, was die Ergebnisse der Studie weiter validiert.



a) b)

Abbildung 10: a) Interfaceelemente des Spiels "Möbelproduktion" für die Post-hoc Erklärungsstudie b) Beschreibungsqualität der Bausteine für die Instanzen der Validierungsstudie (1.0 ist eine perfekte Beschreibung).

In einer weiteren Studie, wurden mithilfe dieses Spiels Klickdaten von 167 Versuchspersonen aufgenommen, um zu untersuchen wie gut die generalisierten Erklärungsbausteine, die in der ersten Studie abgeleitet wurden, die menschlichen Handlungen beschreiben können [p1]. Die Erklärungsbausteine konnten hingehend ihrer Beschreibungsqualität validiert werden: im Durchschnitt konnten 87% der jeweiligen Entscheidungsschritte in einem Trial mit ihnen beschrieben werden (siehe Abbildung 10 b). Die abgeleiteten Erklärungsbausteine bieten also eine gute Grundlage für kognitiv adäquate Strategien, da sie zum einen direkt an die Erklärungen der Menschen angelehnt sind und deswegen der menschlichen Repräsentation entsprechen und zum anderen auch die menschlichen Daten beschreiben.

Um Lösungen für lineare Programme kognitiv adäquat erklären zu können wurden die aus den Verhaltensstudien abgeleiteten Erkenntnisse in einem Prototyp für logische Beschreibungen von Lösungen eingesetzt. Hierfür wurden die Bausteine in einer kontextfreien Grammatik formuliert und mithilfe eines Suchalgorithmus die kürzesten Kombinationen der Bausteine gesucht, die Lösungssequenzen am besten beschreiben. Da nicht alle linearen Programme eine intrinsische Reihenfolge der Teile einer Lösung kodieren, wurden Heuristiken mithilfe der prominentesten menschlichen Strategien bestimmt mit denen die Lösungssequenzen generiert werden können.

Dieser Ansatz zur Lösungsbeschreibung bildet also zum einen die menschliche Repräsentation mithilfe der Grammatik ab, zum anderen wird eine Beschreibung in bedeutsamer Reihenfolge für Menschen generiert, was die Erklärung nachvollziehbarer macht. Der Prototyp wurde auf Instanzen der Möbelfabrik sowie auf einem kleinen Energiebeispiel getestet (Abbildung 11).

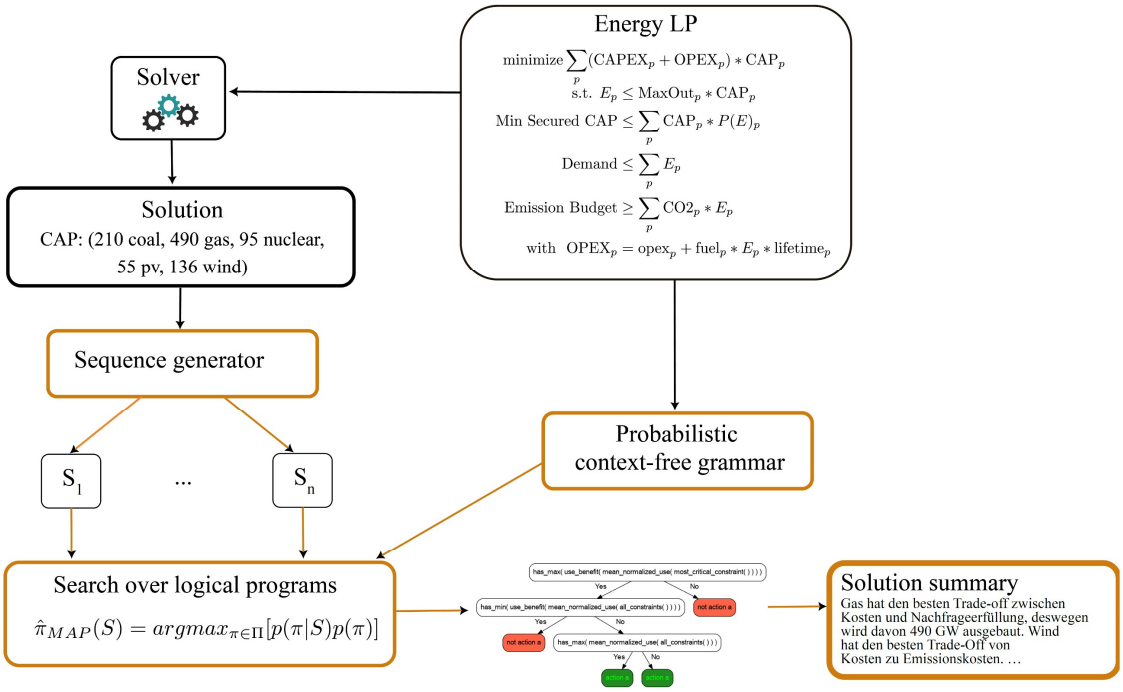


Abbildung 11: Algorithmus zur Suche nach logischen Beschreibungen von Lösungen am Beispiel eines kleinen LP in der Domäne des Energieausbauplanung

AP 4 Extraktion von Wirkzusammenhängen

Das Einsatzgebiet von Energiesystemmodellen umfasst den täglichen Einsatzplan von Kraftwerken bis hin zur langfristigen Ausbauplanung neuer Kraftwerkskapazitäten. Energiesystemmodelle können zur Abbildung von einzelnen Konsumenten/Produzenten bis hin zu internationalen Energienetzen dienen. Entsprechend divers sind auch die zugrundeliegenden mathematischen Modellierungen. Sie reichen von einfachen über sehr umfangreiche lineare Programme bis zu den generell eher opaken Modellen des maschinellen Lernens.

Dank ihres relativ simplen Aufbaus werden lineare Programme häufig bei großen Energiesystemmodellen wie z. B. Ländermodellen verwendet. Die hierbei entstehenden Modelle haben jedoch nicht selten mehrere Millionen Parameter, was die Erklärung der Ergebnisse und des Verhaltens bei der Änderung von Parametern selbst für Experten oftmals sehr herausfordernd oder gar unmöglich macht. [1]

Häufig sind an der Erstellung solcher Energiesystemmodelle unterschiedliche Interessengruppen beteiligt, was zusätzlich den Spielraum für Manipulation öffnet [A7]. Hinzu kommt die Tatsache, dass die eigentliche Entscheidung über die Umsetzung der durch die Modelle errechneten optimalen Handlungsstrategien oft von Laien im Gebiet der Energiesystemmodellierung entschieden werden, die zumeist ein hohes persönliches Risiko tragen, falls sich ihre Entscheidung als schlecht herausstellt. Diese Entscheider können z. B. Politikern, Bürgerversammlungen oder Unternehmensvorstände sein. Um Entscheider bei der Überprüfung der optimalen Handlungsstrategien der Energiesystemmodelle zu unterstützen bedarf es Erklärungen.

Häufig wird im Bereich der Energiesystemmodelle die Sensitivitätsanalyse verwendet, um Erklärungen zu generieren. Eine Sensitivitätsanalyse beschreibt, wie sich Modellergebnisse relativ zu den Änderungen der Eingabegrößen ändern. Die erzeugten Erklärungen

sind somit an die Eingabeparameter eins Modells gebunden und wachsen mit einer steigenden Anzahl an Parametern in ihrem Umfang. Zusätzlich leidet Aussagekraft einer Erklärung, wenn sich der Einfluss auf einen Parameter auf verschiedene Eingabeparameter aufteilt (z. B. bei Zeitreihen).

Um diese Schwächen der Sensitivitätsanalyse bei der Erklärbarkeit von Energiesystem-modellen zu überwinden, wurde eine lokale Erklärungsmethode basierend auf der Methode LIME entwickelt und veröffentlich [10]. Die LIME[2] Methode wurde ursprünglich für die Erklärung von Klassifizierungsalgorithmen im Bereich des maschinellen Lernens entwickelt und setzt auf eine interpretierbare Abstraktionsebene, die es erlaubt Parameter in einer Erklärung zu nutzen, die nicht auf die Eingabeparameter des erklärten Modelles beschränkt und für den menschlichen Nutzer besser verständlich sind. Beispielsweise lassen sich so Konzepte wie Wolken in einer Erklärung nutzen, indem diese auf die eigentlichen Inputs des Energiesystemmodells (Photovoltaikzeitreihe) projiziert werden.

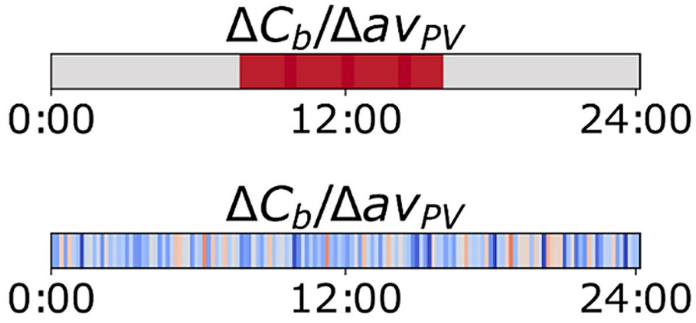


Abbildung 12: Beispiel der Sensitivität einer Batteriekapazität (Cb) relativ zur Änderung an der PV Verfügbarkeit (avPV) - sobald Unsicherheit vorhanden ist (unten) verliert die Sensitivitätsanalyse ihre Aussagekraft.

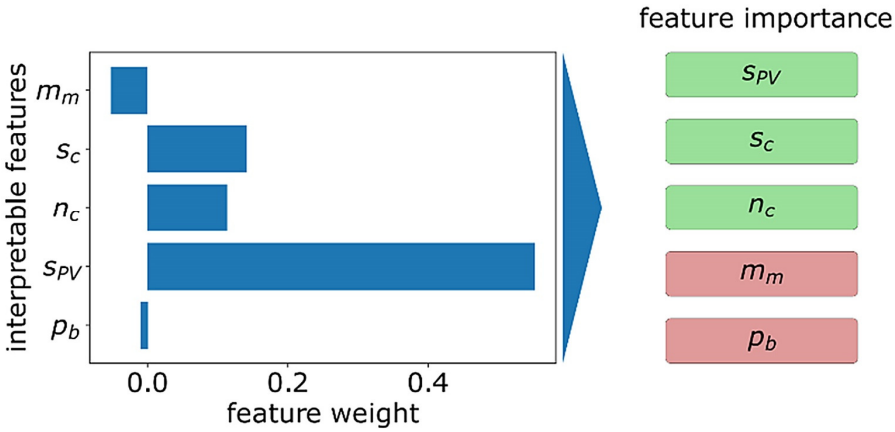


Abbildung 13: Erklärmethode basierend auf LIME - Statt einzelner Zeitschritte werden abstrakte Konzepte wie Wolkengröße (sc) oder Wolkenanzahl (nc) zur Erklärung genutzt. Eine LASSO-Regression[3] erlaubt es die Anzahl von Parametern in der Erklärung und somit ihre Komplexität zu beschränken.

Mit dem entwickelten Ansatz wurde beispielhaft eine Erklärung für Handlungsstrategien in einem Energiesystemmodell von Deutschland erzeugt. Die Restriktionen des verwendeten Energiesystemmodells erzwingen eine CO2 Emissionsreduktion von 80% bis 2050. Das Modell liefert hierfür einen möglichen Transitionspfad mit minimalen Gesamtkosten. Um das das Verhalten des Energiesystemmodells auf den möglichen

[2] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

[3] Tibshirani, Robert (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 58.1 (1996): 267-288.

Transitionspfaden zu erklären, wurden die Preise von fossilen Rohstoffen sowie die Gleichzeitigkeit der Verfügbarkeit von Wind oder Sonne zum Wärmebedarf als abstrakte Parameter benutzt.

Mit der vorgestellten Methode konnte eine Erklärung erzeugt werden, die aufzeigt, dass die zeitliche Divergenz der Erzeugung von erneuerbaren Energien und des Wärmebedarfs nicht dazu führt, dass weniger erneuerbare Erzeugung stattfindet. Die Erzeugung aus erneuerbaren Energien wird stattdessen genutzt um Elektrofahrzeuge zu laden statt Wärmepumpen zu betreiben.

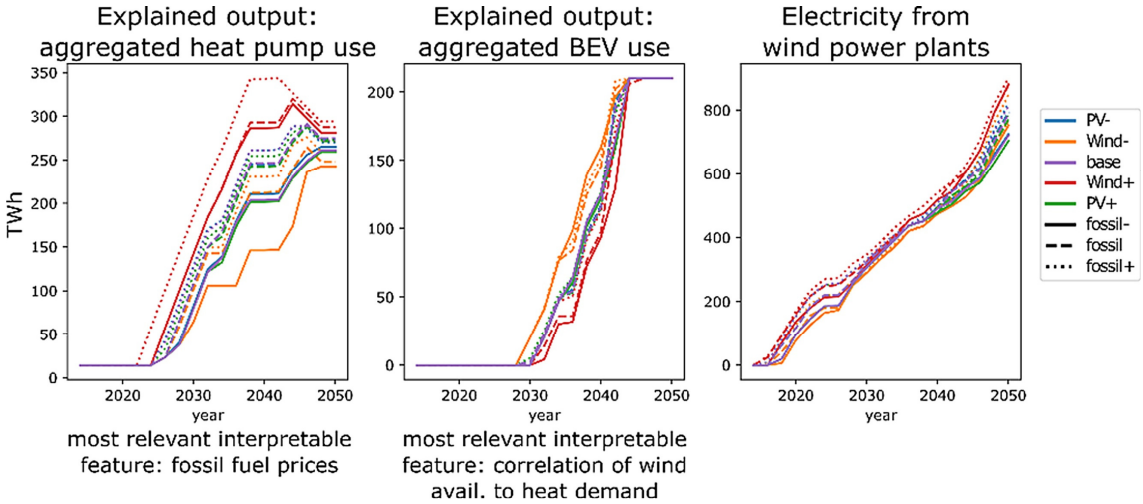


Abbildung 14: Transformation eines Energiesystems in einem Energiesystemmodell von Deutschland. Verschiedene Variationen führen dazu, dass Wärmepumpen oder Elektroautos früher genutzt werden oder später.

Auch wenn die Nutzung der vorgestellten Methode nach wie vor Expertenwissen benötigt, um auf ein Energiesystem angewendet werden zu können, so wurde in [3] bereits gezeigt, wie eine solche Interaktion mit Hilfe eines Chatbots in Zukunft automatisiert werden könnte.

Die praxisrelevanten Ergebnisse und Zwischenergebnisse unseres Projektes wurden insbesondere im Energiebereich mit potenziellen Anwendern, den Industriepartnern Siemens und Entega, projektbegleitend reflektiert.

Generische Übertragbarkeit auf andere Anwendungsszenarien

Die generische Übertragbarkeit der im Projekt erarbeiteten Ansätze auf weitere Anwendungsszenarien ist generell gegeben. Es wurden im Projekt keine Erklärungs-verfahren entwickelt oder verwendet, welche spezifisch für den Energiesektor sind und eine Übertragung auf andere Domänen beeinträchtigen würden. Tatsächlich wurden zudem auch Beispiele aus den Bereichen Gesundheit (u. a. Ernährungsoptimierung und medizinische Diagnose) in AP 1 und AP 2 sowie der Prozesssteuerung und der Logistik in AP 1 und AP 3 (Steuerung einer Möbelproduktion und Routenplanung bei Warenlieferungen) in konkreten Beispielanwendungen untersucht.

AP 5	Projektkoordination und Abstimmung
	Im Projektverlauf wurden regelmäßige Projektmeetings aller Beteiligten auf unterschiedlichen Ebenen des Projektes durchgeführt. Neben einer Vielzahl von Arbeitstreffen in kleineren Kreisen waren dies insbesondere 12 Abstimmungen zum Projektstand und zum Risikomanagement mit allen Forschenden. Dazu kamen formale Treffen aller Beteiligten zumeist inkl. der Industriepartner und, so möglich, dem Projektträger: • 10.06.2020 Projekt-Kickoff • 28.10.2020 Meilenstein 1, 02.11.2021 Meilenstein 2, 31.03.2023 Meilenstein 3 • 17.07.2023 Meilenstein 4, abschließende Ergebnis-Diskussion • 13.10.2023 Projektnachbetrachtung Programmbegleitende Rahmenveranstaltungen des Fördergebers zur Förderlinie fanden während der Projektlaufzeit nicht statt. Jedoch hat PlexPlain am virtuellen „All-Hands-Meeting“ aller BMBF geförderten KI Projekte am 10.01.2022 teilgenommen. Aufgrund der Corona-Pandemie wurde der Großteil der formalen Treffen (insbesondere in 2020 und 2021) virtuell abgehalten. Konferenzreisen der Forschenden waren durch Corona stark eingeschränkt. Versuche mit Testpersonen ließen sich pandemie-konform z. T. virtuell, z. T. unter Beachtung der Hygiene-Regeln umsetzen. Der Anteil an physischen Laborversuchen war jedoch insgesamt geringer als geplant. Im Jahresverlauf 2022 wurden aufgrund der letzten Forschungsergebnisse sowie intensivierter Arbeiten mit Proband:innen und deren Auswertung (im Aufholen der während Corona reduzierten Experimente) in verschiedenen Arbeitspaketen Verzögerungen gegenüber der Projektplanung manifest. Um dieser Situation zu begegnen fanden zunächst monatliche Lenkungstreffen bis Ende 2022 statt. Im Zuge der Reorganisation der Projektrestlaufzeit, wurden die bis dahin im Projekt akkumulierten pandemiebedingten Minderabrufe mit Billigung des BMBF genutzt für eine mittelneutrale Projektlaufzeitverlängerung bis zum 31.07.2023. Durch eine Neuplanung die genannte Verschiebung des Projektendes um 4 Monate und des Meilensteins 3 um 6 Monate (auf den 31.03.2023) konnten die verzögerungsbedingten Risiken minimiert werden und die weitestgehende Erreichung der Projektziele ermöglicht werden.

2.2 Die wichtigsten Positionen des zahlenmäßigen Nachweises

Das Budget des Projekts deckt sich zum allergrößten Anteil mit Personalkosten. Von den zu Projektende nachgewiesenen Kosten entfielen 95% auf wissenschaftliche Mitarbeiter:innen, 2,5% auf wissenschaftliche Hilfskräfte. 1,5% entfielen zudem auf Investitionskosten für Deep Learning High Performance Workstations. Aufgrund der Corona-Pandemie konnte ein Großteil der geplanten Reisen, wie auch Vor-Ort-Experimente mit Versuchspersonen nicht oder nur eingeschränkt stattfinden, weshalb bei Projektende in Summe nur 1% des eingesetzten Budgets auf diese Posten entfielen. Die im Projektverlauf durch Minderabrufe zunächst nicht verwendeten Mittel wurden für das Projektpersonal eingesetzt in Form einer 4-monatigen, kostenneutralen Verlängerung der Projektlaufzeit. Die Verwendung der Fördermittel ist durch die bearbeiteten Projektthemen und die erzielten Ergebnisse dargelegt.

2.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

PlexPlain umfasste anwendungsrelevante Forschungsfragen, für die starkes grundlagenorientiertes Interesse an der TU Darmstadt besteht und ausgewiesene Expertise vorlag und vorliegt. Gleichzeitig war es in seinem Themenumfang notwendigerweise interdisziplinär und bezog Beteiligte aus der Informatik, der Kognitionswissenschaft, der Elektrotechnik sowie assoziiert angebundene Industriepartner mit ein.

Die Organisation eines solch komplexen Forschungsvorhabens im Rahmen der Stellengrundausstattung universitärer Fachgruppen ist in diesem Umfang kaum möglich und bedürfte einer mehrjährigen Vorbereitungsphase um die Verfügbarkeit der jeweiligen Planstellen aufeinander abzugleichen. Eine direkte Auftragsforschung war aus Kosten-Nutzen-Perspektive möglicher Wirtschaftspartner nicht darstellbar aufgrund des Risikos der notwendig hohen Anteile im Bereich der Grundlagenforschung sowie dem geplanten interdisziplinären Gesamt-Projektumfang.

Die Projektförderung durch das BMBF hat daher überhaupt erst ermöglicht die beschriebenen Fragestellungen in dieser Breite zeitnah und sowohl wissenschaftlich fundiert, als auch dennoch durch die Praxis begleitet untersuchen zu können. Die erarbeiteten und hier skizzierten Lösungsansätze haben die dargestellte grundlagenorientierte und interdisziplinäre Forschungsarbeit erfordert. Der explorative Gesamtansatz zwischen notwendiger Grundlagenforschung und Orientierung am gewählten Einsatzfeld und deren Fragestellungen, begründet dabei zugleich die dargestellten Arbeitsaufwände als notwendig im Sinne der Zielerreichung und rechtfertigt die risikoorientierte öffentliche Forschungsförderung.

2.4 Darstellung des voraussichtlichen Nutzens

Eines der Projektziele in PlexPlain waren für Laien und Entwickler verständliche(re) Erklärungen komplexer Planungsszenarien. Hier zeigen die Projektarbeiten Lösungswege auf, wie durch lineare Programme gefundene Lösungen post-hoc erklärt werden können und so eine Vermittlung der komplexen Planungszusammenhänge wirksam unterstützt werden kann. Die Übertragbarkeit dieser Ansätze in andere Domänen als der des Energiesektors ist dabei gegeben und wurde im Projekt auch in weiteren Beispielanwendungen betrachtet.

So handelt es sich bei der Einsatzplanung von verschiedenen Kraftwerken (Kohle, Gas, usw.) in Energiesystemen um eine Problem, in dem Ressourcen möglichst effizient eingesetzt werden sollen, aber gleichzeitig bestimmte Anforderungen erfüllt werden müssen (der Stromverbrauch muss gedeckt sein und CO2-Ziele sollen eingehalten werden). Die Menüplanung in öffentlichen Kantinen ist ein analoges Problem, in dem verschiedene Gerichte zum einen den Bedarf an verschiedenen Nährwerten abdecken müssen, aber gleichzeitig z. B. Oberschranken für den Zuckergehalt bei möglichst geringen Kosten einhalten sollen. Vergleichbare Probleme sind extrem weit verbreitet in Anwendungen. Gestützt durch diese Untersuchungen, erscheint die Eignung unserer Ansätze für eine Nutzung in realen Verwertungskontexten plausibel. Ein konkreter Kandidat ist dabei das im Projekt entwickelte Werkzeug SimplifEx. Im einfachsten Fall erklärt SimplifEx, warum eine Ressource immer besser ist als eine andere, und deshalb zuerst eingesetzt werden sollte, bis sie aufgebraucht ist. In realen Problemen sind Ressourcen aber oftmals in einer Hinsicht gut und in anderer schlecht (z. B. ist Kohle billig, aber produziert viel CO2). SimplifEx kann auch für solche komplexeren Fälle oftmals plausibel machen, wie aufgrund der einzuhaltenden Kriterien die Abwägung zwischen den verschiedenen Ressourcen getroffen werden sollte, um eine möglichst optimale Lösung zu erhalten. Ähnliches gilt für die LIME-basierte Erweiterung der klassischen Sensitivitätsanalyse, die in PlexPlain entwickelt wurde: Sie erweitert bekannte einfache Fälle auf realistischere Fälle, in denen komplexere Abwägungen mit mehreren Variablen vorgenommen werden müssen. Auch diese Analyse wird voraussichtlich einen Nutzen über ihre Anwendung auf Energiesysteme hinaus haben.

Die TU Darmstadt verfolgt mit PlexPlain als universitärer Forschungspartner allerdings keine unmittelbaren, eigenständigen kommerziellen Ziele. Aus wissenschaftlicher Sicht haben die im Projekt entstandenen Publikationen eine sehr gute Grundlage für weitere Forschungen in der wissenschaftlichen Community ermöglicht, als natürlich insbesondere auch für die beteiligten Partner, welche nach Projektende diese Arbeiten in weitere Kontexte übertragen werden.

Die Anschlussfähigkeit der PlexPlain-Themen und -Ergebnisse ist dabei sowohl im wissenschaftlichen, als auch im anwendungsbezogenen Kontext sehr hoch. Aus allgemeiner Perspektive verstärkt die gesellschaftlich stark gestiegene Sichtbarkeit von z. T. kritisch hinterfragten Anwendungen der KI wie Large Language Models (z. B. ChatGPT) das Verständnis für Fragestellungen der Erklärbarkeit von KI-unterstützten Entscheidungen oder KI-erzeugten Ergebnissen. Damit ist auch mittel- bis längerfristig von einer hohen wissenschaftlichen und wirtschaftlichen Relevanz der Erforschung entsprechender Verfahren auszugehen.

2.5 Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Im Rahmen des wissenschaftlichen Projektvorgehens wurden internationale Forschungsansätze und Forschungsergebnisse, welche die Themengebiete von PlexPlain betrafen, verfolgt und in unserer Arbeit und den wissenschaftlichen Publikationen berücksichtigt. Die Referenzen in den unter Abschnitt 2.6 aufgeführten Veröffentlichungen geben daher auch ein Bild der FuE-Ergebnisse Dritter, welche für unsere Arbeit wichtig waren.

Insgesamt war der Projektzeitraum, insbesondere zum Ende hin, eine turbulente Zeit in der KI-Forschung und geprägt durch den rasanten Fortschritt von sogenannten Large-Language-Models (LLMs, wie ChatGPT), der zum Zeitpunkt der Antragstellung noch nicht abzusehen war. Das PlexPlain-Projekt verfolgte einen Ansatz, in dem klassische Methoden und die zum Antragszeitpunkt aktuellen Methoden im Bereich Explainable AI weiterentwickelt werden sollten und zum ersten Mal auf lineare Programme angewandt wurden. Die Entwicklung neuer und innovativer Methoden für lineare Programme basierte dabei auf den Erklärungen von Menschen, die in Verhaltensstudien gewonnen wurden. Diese Erklärungen sind selbstverständlich sprachlich gegeben und die Analyse der Erklärungen erforderte aufwendige qualitative Textanalysen. Ein gewisser Teil des Entwicklungsaufwandes im Projekt war auch, automatisch generierte Erklärungen in natürliche Sprache zu übersetzen. Man könnte nun z. B. von uns gesammelte Erklärungen nutzen, um damit aktuellen LLMs Beispiele für kognitiv adäquate Erklärungen zu liefern, mit dem Ziel bessere sprachliche Erklärungen automatisch generieren zu können. Die Generierung von Erklärungen hat sich in der Zwischenzeit als eine der wichtigsten Anwendungen von LLMs gezeigt und Prompting-Methoden wie Chain-of-Thought oder Tree-of-Thought sind dabei sehr vielversprechend. Bisher hat noch niemand diese Methoden ernsthaft im Bereich Explainable AI eingesetzt, aber es besteht kein Zweifel, dass zukünftige Projekte, die so wie unser Projekt kognitiv adäquate Erklärungen automatisch generieren lassen wollen, an Sprachmodellen nicht vorbeikommen werden.

2.6 Erfolgte oder geplante Veröffentlichung des Ergebnisses

Konferenz- und Journalbeiträge

[1] Jonas Hülsmann and Florian Steinke (2020): Explaining Complex Energy Systems: A Challenge. Poster presented at: Tackling Climate Change with Machine Learning - NeurIPS; December 11, 2020.

[2] Matej Zečević, Devendra Singh Dhami, Athresh Karanam, Sriraam Natarajan, Kristian Kersting (2021): Interventional Sum-Product Networks: Causal Inference with Tractable Probabilistic Models. Published in Proceedings of Neural Information Processing Systems 34.

[3] Jonas Hülsmann, Lennart J. Sieben, Mohsen Mesgar, Florian Steinke (2021): A Natural Language Interface for an Energy System Model. 2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), 2021, pp. 1-5, doi: 10.1109/ISGTEurope52324.2021.9640196.

[4] Matej Zečević, Devendra Singh Dhami, Kristian Kersting (2022). Intriguing Parameters of Structural Causal Models. arXiv preprint arXiv:2105.12697 (v4), under review at IJCAI 2022.

[5] Matej Zečević, Devendra Singh Dhami, Constantin Rothkopf, Kristian Kersting (2022). Structural Causal Interpretation Theorem. arXiv preprint arXiv:2110.02395, under review at ICML 2022.

[6] David Steinmann, Matej Zečević, Devendra Singh Dhami, Kristian Kersting (2022). Machines Explaining Linear Programs. Under review at ICML 2022.

[7] Florian Peter Busch, Matej Zečević, Devendra Singh Dhami, Kristian Kersting (2022). Attributions Beyond Neural Networks: The Linear Program Case. Under review at ICML 2022.

[8] Matej Zečević, Devendra Singh Dhami, Kristian Kersting (2023). Interventions in Graph Neural Networks Lead to New Neural Causal Models. arXiv preprint arXiv:2109.04173, under review at TMLR.

[9] Claire Ott, Inga Ibs, Constantin Rothkopf, Frank Jäkel (2022). Leveraging Human Optimization Strategies for Explainable AI. Talk at Workshop on Human Behavioral Aspects of (X)AI, 2022

[10] Hülsmann, Jonas, Julia Barbosa, and Florian Steinke. 2023. "Local Interpretable Explanations of Energy System Designs" Energies 16, no. 5: 2161. https://doi.org/10.3390/en16052161

[11] Matej Zečević, Florian Peter Busch, Devendra Singh Dhami, Kristian Kersting (2022). Finding Structure and Causality in Linear Programs. Published in International Conference on Learning Representations Workshop on “Objects, Structure and Causality”.

Abschlussarbeiten

[A1] Frodl, E. (2021): The Furniture Company: Building Games to Measure Human Performance in Optimization Problems. Bachelor’s Thesis, Advisors: F. Jäkel, C. Ott. Technische Universität Darmstadt, 2021.

[A2] Sieben, L. (2021): Natural Language Interface for an Energy System Design Tool. Master’s Thesis, Advisors: F. Steinke, J. Hülsmann. Technische Universität Darmstadt, 2021.

[A3] Seng, J. (2021): Causal Discovery in Energy System Models. Master’s Thesis, Advisors: F. Steinke, K. Kersting. Technische Universität Darmstadt, 2021.

[A4] Busch, F. P. (2022): Explaining Neural Network Representations of Linear Programs. Master’s Thesis, Advisors: Kersting, M. Zečević. Technische Universität Darmstadt, 2022.

[A5] Steinmann, D. (2022): Explaining Linear Programs via Neural Attribution Methods. Master’s Thesis, Advisors: Kersting, M. Zečević. Technische Universität Darmstadt, 2022.

[A6] Dotterer, S. (2022) Investigating the Influence of Different Cost-Profit Ratios on Human Performance and Strategies in Optimization Problems in an Eye Tracking Experiment. Bachelor’s Thesis, Advisors: Rothkopf, C. A., Ibs, I. Technische Universität Darmstadt, 2022

[A7] Uetz, P. (2022): Investigation of the Vulnerability of Energy System Models to Adversarial Attacks. Master’s Thesis, Advisors: F. Steinke, J. Hülsmann. Technische Universität Darmstadt, 2022.

[A8] Pohl. A (2023): Die Heidelberger Struktur-Lege-Technik als Werkzeug zur Analyse subjektiver Theorien im Kontext des Planspiels Energiewende. Bachelor’s Thesis, Advisors: F. Jäkel, C. Ott. Technische Universität Darmstadt, 2023

[A9] Rödling, S. (2023): Providing Causal Explanations Over Time: An Extension of SCE for Time-Series Data. Master’s Thesis, Advisors: Kersting, M. Zečević. Technische Universität Darmstadt, 2023.

Preprints und angenommene Konferenz- bzw. Journalbeiträge

[p1] Inga Ibs, Claire Ott, Frank Jäkel, Constantin Rothkopf, (under review). From human explanations to explainable AI: Insights from constrained optimization.

[p2] Claire Ott and Frank Jäkel, (under review). SimplifEx: Simplifying and explaining linear programs.

[p3] Claire Ott, Inga Ibs, Constantin Rothkopf, Frank Jäkel, (in prep). Unveiling the relationship between tasks: Optimization as a case for taxonomic analysis.

[p4] Matej Zečević, Devendra Singh Dhami, Constantin Rothkopf, Kristian Kersting (2022). Causal Explanations of Structural Causal Models. arXiv preprint arXiv:2110.02395.

[p5] David Steinmann, Matej Zečević, Devendra Singh Dhami, Kristian Kersting (2022). Machines Explaining Linear Programs. arXiv preprint arXiv:2206.07194.

[p6] Florian Peter Busch, Matej Zečević, Devendra Singh Dhami, Kristian Kersting (2022). Attributions Beyond Neural Networks: The Linear Program Case. arXiv preprint arXiv: 2206.07203.

[p7] Matej Zečević, Devendra Singh Dhami, Kristian Kersting (2023). Interventions in Graph Neural Networks Lead to New Neural Causal Models. arXiv preprint arXiv:2109.04173.