CALIBRaiTE

Reliability Displays für die Vertrauenskalibrierung KI-basierter Systeme

Programm / Ausschreibung	Ideen Lab 4.0, Ideen Lab 4.0, Ideen Lab4.0 - Ausschreibung 2019	**Status**	abgeschlossen
Projektstart	01.02.2020	**Projektende**	31.03.2021
Zeitraum	2020 - 2021	**Projektlaufzeit**	14 Monate
Keywords	Vertrauen, KI, künstliche Intelligenz, Mensch-Maschine Interaktion, Akzeptanz, User Experience		

Projektbeschreibung

Kurzbeschreibung

Das Vertrauen in ein automatisiertes System kennzeichnet sich durch die Erwartungshaltung, dass dieses eine Person in einer Situation unterstützt, die durch Ungewissheit und Verletzbarkeit gekennzeichnet ist. Wichtig ist es somit zu wissen, in welcher Situation man sich auf eine intelligente Funktion verlassen sollte und wann nicht. Wenn die Verlässlichkeit der intelligenten Funktion unter- oder überschätzt wird, sie also nicht gut genug “kalibriert” ist, dann führt dies zu Distrust oder Overtrust. Wenn diese Phänomene häufig vorkommen, kann sich dies negativ auf die langfristige Akzeptanz von KI-basierten Anwendungen auswirken. Reliability Displays sind in den letzten Jahren vorgeschlagen worden, um solche Informationen zur Verlässlichkeit von intelligenten Systemfunktionen zu bieten und somit die Erwartungshaltung mit den eigentlichen Systemfähigkeiten in Einklang zu bringen. Es handelt sich bei Reliability Displays keineswegs um den Versuch, "NutzerInnen zu kalibrieren" bzw zu bestimmten Verhaltensweisen zu bringen, sondern es soll diesen eine Möglichkeit an die Hand gegeben werden, ihre eigene akzeptanz- und vertrauensbezogene Einstellung zu einer Systemfunktion anzupassen. Es erscheint einleuchtend, dass Informationen zur Verlässlichkeit (Reliability Displays) eine wichtige Rolle spielen können. Daher ist es umso erstaunlicher, dass diese weder weit verbreitet sind noch genau erforscht sind. Ein wichtiger Beitrag des Sondierungsprojekts CALIBRaiTE ist es, Reliability Displays in den Vordergrund der Betrachtung zu rücken und deren Potentiale und Einschränkungen sehr sichtbar zu präsentieren und zu reflektieren. Durch diese vorbereitende Exploration könnte dann die Grundlage für effektive Forschungs- und Entwicklungsentscheidungen getroffen werden.

Initiativen für eine systematischere Betrachtung von Reliability Displays sind in den letzten Jahren im Bereich des (teil-)automatisierten Fahrens unternommen worden, wie beispielsweise Spurhalteassistenten. Die Ergebnisse erster Studien in diesem Bereich sind ermutigend, allerdings sind hier noch weitere Untersuchungen notwendig. In anderen Systembereichen mit intelligenten Funktionen, insbesondere der industriellen Produktion, gibt es zwar in den letzten Jahren ein größeres Bewusstsein für nutzerInnenzentriertes Design - allerdings entstehen Designs eher arbiträr in Form von informellen praktischen Belangen geleitet, wodurch bisher Reliability Displays in diesem Sktor noch nicht übergriefend analysiert worden sind. Eine grundlegende Herausforderungen lautet in diesem Sinne: wie können Reliability Displays in Zusammenhang mit KI-basierten Systemen verwendet werden, sodass ein angemessenes Vertrauensniveau ermöglicht wird?

Ein in letzter Zeit besonders relevant gewordener Bereich innerhalb der industriellen Produktion, der sich durch ein hohes

Maß an KI-basierten Ansätzen auszeichnet, ist die vorausschauende Wartung (predictive maintenance). Dieser Geschäftsbereich adressiert einen wichtigen Hoffnungsmarkt für österreichische Unternehmen und sollte somit aktiv in ihren Bestrebungen für verbesserte, KI-basierte Produkte unterstützt werden. Einer dieser innovativen Produktbereiche, die einen Differenzierungsfaktor für österreichische Unternehmen darstellen können sind verbesserte Mensch-Maschine Schnittstellen. Auch hier sollten Nutzer ständig im Bilde darüber sein, ob die Vorhersagen solcher Systeme vertrauenswürdig sind und ob sie deren Empfehlungen befolgen sollen oder nicht. Der heutige betriebliche Alltag zeigt allerdings ein gegensätzliches Bild: derzeitige Dashboards an Produktionsstandorten lassen NutzerInnen im Unklaren darüber, auf welcher Datenqualität ihre Vorschläge und Bewertungen beruhen und was die zugrundeliegende Algorithmik ist.
Im konkret betrachteten Fall werden bei einer Assembly Line eines Automobilherstellers, welche die Aufgabenstellung einer Just-in-Sequence-Produktion hat, wo zu verbauende Teile nicht zwischengelagert werden sondern direkt an die Linie angeliefert werden. Eine Verzögerung mehrer Teile macht eine sehr teure Neuplanung bzw ein Rescheduling notwendig oder führt im schlimmsten Fall zu einem Linienstop. In der prädiktiven Wartung wird daher das Wartungsintervall sowie die Wahrscheinlichkeit für einen wartungsbedingten Ausfall für jede Maschine berechnet und in die Produktionsplanung berücksichtigt. Das Ziel ist es, durch prädiktive Wartung die vorhersehbaren Wartungsintervalle so zu planen, dass diese den Betrieb nicht stören und gleichzeitig die Wahrscheinlichket eines unvorhersehbaren wartungsbedingten Ausfalls möglichst auf Null reduziert. Um sowohl den Ist-Zustand der Maschine als auch den zu erwartenden zukünftigen Zustand der Maschine anzuzeigen, werden aufgrund von verschiedener Sensordaten an der Maschine sowie der Maschine selbst Daten gesammelt, diese werden aggregiert und erlauben dadurch eine Darstellung des Zustandes sowie eine Abschätzung über die Wahrscheinlichkeit eines Ausfalls in der Zukunft.
Eine dringende Herausforderung aus diesem speziellen österreichischen Industriebereich ist es, die tatsächlichen Daten sowie die errechneten Daten zur zukünftigen Einschätzung so darzustellen, dass es zu keiner Missinterpretation kommt. Weiters können bereits von einer vertrauenswürdigen Person geschätzte Zustände vertrauensvoller dargestellt werden als Daten die ohne menschlicher Kontrolle aggregiert wurden. Weiters besteht aus diesem Anwendungsbereich der Bedarf, Ergebnisse einer sub-symbolischen KI - die eine gewisse Abstraktion und Ungenauigkeit aufgrund ihre Struktur haben muss - anders darzustellen als beispielsweise Ergebnisse einer symbolischen KI, die als formal korrekt anzusehen ist.
Das Projekt CALIBRaiTE setzt es sich daher zum Ziel, wichtige Weichenstellungen für die Vertrauenskalibrierung zu liefern. Erstens soll eine Abschätzung bzgl der Eignung von Reliability Displays für prädiktive Funktionen in der industriellen Produktion ermöglicht werden, um eine Potentialisierung für weitere entsprechende Vorhaben zu ermöglichen. Hierbei ist insbesondere der Aspekt der individuellen und innerbetrieblichen Akzeptanz relevant. Zweitens soll überprüft werden, inwiefern die Design Pattern Methode ein passender Ansatz für die Entwicklung von Reliability Displays sein sollen. Dies soll anhand der testweisen Erstellung beispielhafter, Reliability Display Designmuster für den untersuchten Anwendungsfall erprobt werden. Um die Wirkung von Reliability Displays einschätzen zu können, wird drittens aus der Informatikperspektive die für Reliabiility Displays zentrale Komponente der Ungewissheit (Uncertainty) in Form von Szenarien, Kontextfaktoren und Key Performance Indicators (KPIs) untersucht, wie zB die zugrundeliegenede Datenqualität, die Relevanz des Ereignisses und die Komplexität der Situation in prädiktiven Wartungsservices.
In Zusammenhang mit allen drei oben genannten Zielen zielt CALIBRaiTE auf die Erprobung eines partizipativen Entwicklungs- und Reflexionsansatzes für Reliability Displays ab, der sich durch eine iterative Herangehensweise auszeichnet, bei der zuerst das Anwendungsszenario der prädiktiven Wartung im Produktionsberiech unter Einbindung verschiedener Stakeholder untersucht wird, und in der Folge mit ausgewählten NutzerInnen des Systems mit entsprechender Domänenexpertise ein co-kreativer Patterndesign Prozess durchgeführt wird, welcher dann in Form eines Vergleichs mehrerer Patternvarianten weitergeführt wird. Das ultimative Ziel des Sondierungsprojekts ist es, Entscheidungen

zu Folgeaktivitäten zu ermöglichen, sowie diese bestmöglich zu motivieren. Neben der grundsätzlichen Abschätzung des Potentials bzw der Eignung kann die Erprobung von der Design Pattern Methodik, die technische Modellierung der Ungewissheit, sowie der partizipative Entwicklungsprozess hierfür einen signifikanten Beitrag leisten.

Abstract

Trust in an automated system is characterized by the expectation that it will support a person in a situation characterized by uncertainty and vulnerability. It is therefore important to know in which situation one should rely on an intelligent function and when not. If the reliability of the intelligent function is underestimated or overestimated, i.e. if it is not "calibrated" well enough, this leads to distrust or overtrust. If these phenomena occur frequently, this can have a negative impact on the long-term acceptance of AI-based applications. Reliability displays have been proposed in recent years to provide such information on the reliability of intelligent system functions and thus to align expectations with actual system capabilities. Reliability displays are by no means an attempt to "calibrate" users or to bring them to certain behaviours, but rather to provide them with an opportunity to adapt their own acceptance- and trust-related attitude to a system function. It seems plausible that reliability displays can play an important role in AI-based systems. It is therefore astonishing that these are neither widespread nor thoroughly researched. An important contribution of the CALIBRaiTE exploratory project is to focus on reliability displays and to present and reflect their potentials and limitations very visibly. This preparatory exploration could then form the basis for effective research and development decisions.

Initiatives for a more systematic consideration of reliability displays have been undertaken in recent years in the field of (partially) automated driving, such as lane departure warning systems. The results of initial studies in this area are encouraging, but further research is needed. In other system areas with intelligent functions, especially in industrial production, there has been a greater awareness of user-centered design in recent years - however, designs tend to emerge arbitrarily in the form of informal practical concerns, so that reliability displays in this sector have not yet been extensively analysed. A fundamental challenge in this sense is: how can reliability displays be used in conjunction with AI-based systems to enable an adequate level of trust?

One area within industrial production that has recently become particularly relevant and is characterized by a high degree of AI-based technological advancement is predictive maintenance. This business area addresses an important market of hope for Austrian companies and should therefore be actively supported in their efforts for improved, AI-based products. One of these innovative product areas, which can represent a differentiation factor for Austrian companies, relates to optimized man-machine interfaces. Here, too, users should be constantly informed as to whether the predictions of such systems are trustworthy and whether they should follow their recommendations or not. However, everyday business life shows a contrasting picture: current dashboards at production sites leave users unclear about the data quality on which their suggestions and evaluations are based and the status of underlying algorithms.

In the concrete case under consideration, an assembly line of an automobile manufacturer, which has the task of a just-in-sequence production, where parts to be installed are not stored temporarily but are delivered directly to the line. A delay of several parts makes a very expensive replanning or rescheduling necessary or, in the worst case, leads to a line stop. In predictive maintenance, the maintenance interval and the probability of a maintenance-related breakdown are therefore calculated for each machine and taken into account in production planning. The aim is to use predictive maintenance to plan the predictable maintenance intervals in such a way that they do not disrupt operation and at the same time reduce the probability of an unforeseeable maintenance-related failure to zero. In order to display both the actual condition of the machine and the expected future condition of the machine, data is collected on the basis of various sensor data on the machine and the machine itself; these data are aggregated and thus allow a representation of the condition as well as an

estimate of the probability of a failure in the future.

Endberichtkurzfassung

Im Projekt CALIBRaiTE wurde erforscht, wie langfristig Vertrauen im Umgang mit intelligenten Systemen im industriellen Kontext aufgebaut werden kann. Dies wird durch ein „ehrliches“ Systemverhalten ermöglicht, bei dem in jeder Nutzungssituation die Verlässlichkeit von Systemfunktionen (mittels sogenannter „Reliability Displays“) kommuniziert wird. In einem partizipativen Prozess mit Expert*innen und Nutzer*innen wurden hierfür die notwendigen Ansätze zur Datenmodellierung, User Interface Patterns und Evaluierungsmethoden erarbeitet.

Am Beispiel des Prozessmanagements von Bauprojekten wurden Design Patterns entwickelt, welche direkt umsetzbare Lösungen für Interfacedesigns von Reliabilitätsindikatoren zur Vertrauenskalibrierung in Decision Support Systems (DSS) im Kontextfeld Building Information Management (BIM) enthalten. Es wurden Design Patterns zu den Faktoren "expertisebasierte Nutzerrollen", "Reliabiltät durch zeitliche Nähe" und "Visualisierung von Pönalen" anhand eines bereits zuvor in anderen Kontexten angewandten und bewährten Vorgehens. Ein jedes Pattern enthält eine detaillierte Beschreibung der Ausgangs- und Problemstellung inklusive des Kontextes, gefolgt von der Designlösung und illustrativen Beispielen aus existierenden Anwendungen. Die Patterns sind anwendungsorientiert und anwendungsnahe, kontextspezifische Lösungen mit hohem Detailgrad, welche mit der in dem jeweiligen Pattern enthaltenen Informationen alleine und ausreichender Domänenexpertise angewendet werden können. Sie stellen daher sowohl konkrete und verwendungsbereite Interface-Designlösungen für drei primäre Herausforderungen der Vertrauenskalibrierung von DSS im BIM-Kontext dar, als auch eine Basis zur Erforschung und Konzeption weiterer Methoden der Vertrauenskalibrierung in DSS.

Weiters wurden Methoden der Datenmodellierung entwickelt, welche die Identifizierung relevanter Key-Indikatoren für die Performance des Systems erlaubte, um eine optimistische, pessimistische und durchschnittliche Vorhersage abzugeben. Damit wir eine ausreichende Datengrundlage für die Untersuchung der Zuverlässigkeit in D4.2 schaffen, wurden Daten für die definierten Szenarien simuliert und ein Kompendium von Testszenarien entwickelt. Für jedes Szenario wurden dazu abhängig von den oben genannten drei Faktoren "expertisebasierte Nutzerrollen", "Reliabiltät durch zeitliche Nähe" und "Visualisierung von Pönalen" jeweils mehrere Varianten für eine optimistische, durchschnittliche und pessimistische Kostenabschätzung auf Basis der zugrundeliegenden KPI-Datenmodelle errechnet. Für die entwickelten KPI-Datenmodelle wurden in einem Dashboard Prototyp verschiedene Arten von Widgets implementiert. Die entwickelte Architektur für das Dashboard wurden dokumentiert und die vier häufigsten Visualisierungen exemplarisch beschrieben. Darüber hinaus wurden verschiedene Möglichkeiten für eine KI-basierte Modellierung der Reliability untersucht, basierend auf Decision Tree Modellen und einem Conditional Inference Tree Ansatz. Beide Methoden erlauben die grafische Darstellung der Entscheidungswege und helfen es, die Ergebnisse als Nutzer besser interpretieren zu können. Für die gewählten Szenarien war die erste Methode ausreichend und lieferte unter den gewählten Voraussetzungen Genauigkeiten um die 90%. Im Falle des Conditional Inference Tree Ansatzes erhielten wir vergleichbare Ergebnisse und rechnen damit, dass diese Methode für komplexere Modelle geeigneter ist, da Korrelationen zwischen verschiedenen Features besser berücksichtigt werden können.

Ein weiterer Beitrag des Projekts war die Erprobung eines partizipativen Entwicklungs- und Reflxionsansatzes für Reliability Displays, der sich durch eine iterative Herangehensweise auszeichnet. Die wichtigsten Iterationen bei diesem partizipativen bzw co-kreativen Entwicklungsprozess umfassten die Untersuchung von Anwendungsszenarien der prädiktiven Wartung im

Produktionsbereich unter Einbindung verschiedener Stakeholder*innen, sowie in der Folge einen co-kreativen Prozess zur Generierung von Patterns und deren Evaluation mit ausgewählten Nutzer*innen des Systems mit entsprechender Domänenexpertise. Eine empirischen Pilotstudie ergab eine großteils positive Wirkung von Reliability Displays. So erhöhten Reliability Displays die Bereitschaft, erste Abschätzungen zu machen, selbst wenn man noch nicht sehr mit dem Projekt und System vertraut ist. In Übereinstimmung mit diesem Befund waren die Anwender (die Hälfte der Teilnehmer dieser Studie) tendenziell zuversichtlicher bei ihren eigenen Entscheidungen, wenn ihnen Reliability Displays gezeigt wurden, während die andere Hälfte indifferent war. Die Verfügbarkeit von Reliability Displays förderte erwartungsgemäß auch die Tendenz, das Vertrauen zu kalibrieren, d. h. ihr Vertrauen in das System gemäß der angezeigten Verlässlichkeit anzupassen.

Weiters ergab die Pilotstudie, dass Reliability Displays positive Effekte auf das Vertrauen und die Akzeptanz in komplexen Prozessmanagement Szenarien wie dem Prozessmanagement von Bauprojekten haben können. Die Ergebnisse zu den Technologieakzeptanzfaktoren zeigen, dass Reliability Displays die wahrgenommene Nützlichkeit eines Prozessmanagementsystems signifikant erhöhen. Weiters deuten die Ergebnisse der dritten Phase der Pilotstudie darauf hin, dass Reliability Displays in Prozessmanagement User Interfaces auf mehreren Ebenen bereitgestellt werden sollten. Die stärkste Präferenz wurde für Erklärungselemente (Reliability Explanations) gefunden, die Hintergrundinformationen über diejenigen Faktoren liefern, die zur Gesamt-Zuverlässigkeitsbewertung beitragen. Darüber hinaus wurden erstmals in einer systematischen empirischen Studie Faktoren zur Ermittlung der Datenzuverlässigkeit untersucht und darauf aufbauend spezifischere Anforderungen an die Gestaltung von Zuverlässigkeitserklärungen abgeleitet.

Ausgehend von den Erfahrungen und Ergebnissen aus dem Projekt wurden Handlungsempfehlungen abgeleitet. Diese beziehen sich einerseits auf die einzelnen untersuchten Aspekte (Analyse des Forschungsstands, die Auswahl von Anwendungsfällen, die Entwicklung von Patterns und Designvorschlägen, die Datenmodellierung und die nutzer*innenzentrierte Evaluierung). Als Ausblick wird auch eine Reihe allgemeiner, strategischer Handlungsempfehlungen für Forschungsprogramme und -projekte vorgestellt.

Projektkoordinator

- AIT Austrian Institute of Technology GmbH

Projektpartner

- BOC Asset Management GmbH
- Universität Salzburg