Stephan Wronkowsk-Elster

ABSTRACT

Various decisions relating to critical areas, such as medicine, finance, security, and defence, are being passed on to Artificial Intelligence (AI) algorithms with increasing regularity. Therefore, it is of immense importance to understand why an artificial intelligence model makes a particular decision or performs a specific action. This so-called explainability means that developers, users, and society remain able to comprehend -- as well as trust – certain judgements or results of activities; this, in turn, will allow for better decisions in cooperation between man and machine, or among autonomous machines. The research investigating these aspects is known as Explainable Artificial Intelligence – (short XAI1 ). Strategy, corporate planning, and decision-making are core managerial functions. IT systems and approaches like Business Analytics, Predictive and Prescriptive Analytics have increasingly supported these tasks in the past twenty years. Artificial Intelligence is the next step in developing such systems to support managerial capacities. However, if AI activities cannot be explained with respect to human communication, understandability, and readability, their users (especially the managerial and supervisory bodies) will not trust them (Chiusi et al., 2020; Been Kim et al., 2016; Bejger & Elster, 2019; Bejger & Elster, 2021; Christoph Molnar et al., 2019; Gilpin et al., 2019; de Graaf & Malle, 2017; Gunning, 2016; High-Level Expert Group on Artificial Intelligence, 2019; Lipton, 2017;Marco Tulio Ribeiro et al. 2016; Miller, 2018; Sameer Singh et al., 2019; Wachter et al., 2017; etc.) . AI's decisions and actions must be kept transparent, interpretable, and explainable to be considered trustworthy and reliable. The AI models (AI models, AI agents, and AI systems are terms used synonymously in this work) must also be held accountable for the decisions and actions they undertake (Doshi-Velez et al., 2017; Doshi-Velez & Kim, 2017, Wachter et al., 2017), as a lack of trust may lock substantial untapped potential for the increased growth of this new and promising technology, particularly in the field of integrated Corporate Planning and its objectives. 1 Following Chari et al. (2020), the short term XAI was first coined by DARPA and therefore focus on a specific project. - Explainable AI is much broader, but this thesis is using both the term explainable AI and the abbreviations XAI or xAI. v The process industry (in short PI) plays a significant role in worldwide business, mainly in the pharmaceutical industry; it plays a leading role in protecting against (or even prevention of) deadly pandemics, for instance, as well as in the fight against cancer. Therefore, the PIs are central to transforming raw materials by way of applying suitable systems or methods into finished products. PIs can be classified with regards to their feedstock type or products used, for example, petroleum refining, mineral processing, chemical processing, fertilizers, food, and pharmaceuticals (Brennan, 2020). This work focused primarily on the chemical and pharmaceutical (in short, pharmacy) industry. Corporate planning is the core process in the management cycle and deals with the prediction and achievement of future desired states that deviate from the current state. The importance of this process is particularly relevant for companies that produce with high fixed costs and thus must guarantee capacity utilisation in the future, as their production capacity cannot be flexibly adjusted. This is even more true for companies in the process industry due to their networking in highly complex supply chains up- and downstream (Elster, 2009). Therefore, optimised planning and decision-making, which uses, for example, modern tools such as artificial intelligence for strategic and tactical planning, is of immense importance for these companies. This pertains to chemicals and pharmacy, specifically scenario planning, integrated business planning, and decision-making. The use of AI, especially subsymbolic black-box models, presents the above challenges. Hypothesis: By developing a reference architecture for an explainable AI system that could combine both subsymbolic and symbolic approaches, confidence in AI models and, thus, decision-making in corporate planning can be improved. The primary goal of this dissertation is to establish and create a reference system architecture that promotes explainable artificial intelligence, with the aim of improving decision-

making capabilities to facilitate better business planning within the process industry. The research has resulted in a reference system architecture for trustworthy AI in corporate planning, which is the main contribution of this work. To the author's knowledge, there are no previous or other comparable works in this domain. vi This work examines a crucial research question: How can an explainable artificial intelligence system, or agent, be created and integrated into the planning framework of the process industry to increase trust in decision-making AI systems by improving their transparency and decision quality? The research method is based on Design Science (Hevner et al., 2004; Wieringa, 2014). The reference architecture is supported by relevance and scientific rigour. Hevner et al. (2004) define seven guidelines for understanding, executing and evaluating design-science research. The current research produced the reference system architecture for trustworthy AI in corporate planning in the process industry (Galster, 2011; Nakagawa, 2014). The issue addressed in this research holds significant importance for both theoretical and practical applications, with recent studies supporting its relevance. (Chiusi et al., 2020; Bejger & Elster, 2019; Bejger & Elster, 2020; Molnar, 2019; Singh et al., 2019; Tulio Ribeiro et al., 2016; Been Kim, et al, 2016; Gunning, 2016; High-Level Expert Group on Artificial Intelligence, 2019; Wachter et al., 2017; Chakraborti et al. 2020; Willms & Brandenburg, 2019; etc.) The evaluation of the design was conducted in two stages. Firstly, the design process was assessed based on the design principles outlined by Hevner et al. (2004). Secondly, a survey was conducted with experts in the fields of architecture and corporate planning to evaluate the design further. The survey questions were based on criteria derived from best practices in research and studies. (Bass et al., 2021; Vasconcelos et al., 2005). They were introduced using the method described by Saunders et al. (2023) and Sekaran and Bougie (2019). After conducting a survey, statistical methods were used to analyse the results. The group of experts provided additional concerns, requirements, and constraints identified as gaps in the thesis. These gaps will be considered during the next iteration of the design cycle. The survey results confirm the hypothesis that the developed reference architecture can serve as a viable solution to the stated problem. The design science approach mentioned above was used to design and build a reference architecture called “Re_fish” (s combination of “Rejewski” and “Babelfish”, as a tribute to Marian Rejewski -- the leading Polish scientist who broke the Enigma code, and the vii Babelfish – a fictional entity and universal decoder for any form of language in the universe2 ) that can be used in a (corporate) planning context within the process industry, by using a knowledge-based hybrid approach (Hitzler et al. 2021; Hitzler & Sarker, 2022; Hochreiter, 2022; Niu et al. 2022; Futia & Vetrò, 2020; Marcus, 2020; Marcus & Davis, 2021; Sohrabi et al., 2018; Tiddi, et al., 2020). 2 “The Babelfish,“ said the Hitch Hiker’s Guide to the Galaxy quietly, “is small, yellow and leech-like, and probably the oddest thing in the Universe. It feeds on brainwave energy […]” (Adams, 2010, p. 60) viii ABSTRACT Różne decyzje w krytycznych obszarach, takich jak medycyna, finanse, bezpieczeństwo i obrona narodowa, są coraz częściej przekazywane algorytmom sztucznej inteligencji (AI). Dlatego niezwykle ważne jest zrozumienie, dlaczego model sztucznej inteligencji podejmuje określoną decyzję lub wykonuje określone działanie. Tak zwana wyjaśnialność oznacza, że programiści, użytkownicy i społeczeństwo są w stanie zrozumieć i zaufać pewnym osądom lub wynikom działań, co z kolei umożliwia podejmowanie lepszych decyzji we współpracy człowiek-maszyna lub między autonomicznymi maszynami. Badania zajmujące się tymi aspektami nazywane są Explainable Artificial Intelligence (w skrócie XAI). Strategia, planowanie biznesowe i podejmowanie decyzji to podstawowe funkcje zarządzania. Systemy i podejścia IT, takie jak Business Analytics, Predictive i Prescriptive Analytics, w coraz większym stopniu wspierały te zadania w ciągu ostatnich dwudziestu lat. Sztuczna inteligencja jest kolejnym krokiem w rozwoju takich systemów wspierających możliwości zarządzania. Jeśli jednak działań AI nie można wyjaśnić w kategoriach ludzkiej komunikacji, zrozumiałości i czytelności, ich użytkownicy (zwłaszcza organy zarządzające) nie będą im ufać (Chiusi i in., 2020; Been Kim i in. 2016; Bejger & Elster, 2019; Bejger & Elster, 2021; Christoph Molnar i in, 2016; Bejger & Elster, 2019; Bejger & Elster, 2021; Christoph Molnar et al, 2019; Gilpin et al, 2019; de Graaf & Malle, 2017;

Gunning, 2016; High-Level Expert Group on Artificial Intelligence, 2019; Lipton, 2017; Marco Tulio Ribeiro et al, 2016; Miller, 2018; Sameer Singh et al, 2019; Wachter et al, 2017; etc.). Decyzje i działania sztucznej inteligencji muszą być przejrzyste, możliwe do zinterpretowania i wyjaśnienia, tak aby można je było uznać za godne zaufania i wiarygodne. Modele AI (modele AI, agenci AI i systemy AI są używane w niniejszym dokumencie jako synonimy) muszą również ponosić odpowiedzialność za podejmowane przez siebie decyzje i działania (Doshi-Velez i in., 2017; Doshi-Velez & Kim, 2017, Wachter i in., 2017), ponieważ brak zaufania może zablokować ogromny i niewykorzystany potencjał wzrostu tej nowej i obiecującej technologii, zwłaszcza w dziedzinie biznesu, a zwłaszcza w planowaniu w przedsiębiorstwie. ix Przemysł przetwórczy (w skrócie PI) odgrywa znaczącą rolę w globalnej gospodarce, zwłaszcza w przemyśle farmaceutycznym. Przemysł farmaceutyczny odgrywa wiodącą rolę np. w ochronie przed (lub nawet zapobieganiu) śmiertelnym pandemiom i np. w walce z rakiem. PI mają zatem kluczowe znaczenie dla przekształcania surowców w gotowe produkty poprzez zastosowanie odpowiednich systemów lub metod. PI można sklasyfikować zgodnie z charakterem wykorzystywanych surowców lub produktów, np. rafinacja ropy naftowej, przetwarzanie minerałów, przetwarzanie chemiczne, nawozy, żywność i farmaceutyki (Brennan, 2020). W niniejszej pracy skupiono się przede wszystkim na przemyśle chemicznym i farmaceutycznym (w skrócie farmaceutycznym). Planowanie w przedsiębiorstwie jest podstawowym procesem w cyklu zarządzania, którego zadaniem jest stawianie prognoz i osiąganie przyszłych stanów docelowych, które odbiegają od stanu rzeczywistego. Znaczenie tego procesu jest szczególnie istotne dla firm, które produkują z wysokimi kosztami stałymi, a tym samym muszą zagwarantować wykorzystanie mocy produkcyjnych w przyszłości, ponieważ ich zdolności produkcyjne nie mogą być elastycznie dostosowywane. Dotyczy to szczególnie firm z branży przetwórczej ze względu na ich wzajemne powiązania w wysoce złożonych łańcuchach dostaw wyższego i niższego szczebla (Elster, 2009). Zoptymalizowane planowanie i podejmowanie decyzji, z wykorzystaniem np. nowoczesnych narzędzi, takich jak sztuczna inteligencja do planowania strategicznego i taktycznego, ma zatem ogromne znaczenie dla tych firm. Dotyczy to sektora chemicznego i farmaceutycznego, w szczególności planowania scenariuszy, zintegrowanego planowania i podejmowania decyzji. Dzięki wykorzystaniu sztucznej inteligencji, a zwłaszcza podsymbolicznych modeli czarnej skrzynki, pojawiają się powyższe wyzwania. Hipoteza: Opracowując architekturę referencyjną dla wyjaśnialnego systemu sztucznej inteligencji, który może łączyć zarówno podejście subsymboliczne, jak i symboliczne, można zwiększyć zaufanie do modeli sztucznej inteligencji, a tym samym podejmowanie decyzji w planowaniu biznesowym. Głównym celem niniejszej rozprawy jest opracowanie architektury systemu referencyjnego, która promuje wyjaśnialną sztuczną inteligencję w celu poprawy zdolności decyzyjnych i umożliwienia lepszego planowania przedsiębiorstw w przemyśle przetwórczym. Wynikiem x badań jest architektura systemu referencyjnego dla godnej zaufania sztucznej inteligencji w planowaniu biznesowym, która stanowi główną część niniejszej rozprawy. Zgodnie z wiedzą autora, nie ma wcześniejszych lub porównywalnych prac w tej dziedzinie. W niniejszej rozprawie kluczowym pytaniem badawczym jest: W jaki sposób można stworzyć wytłumaczalny system sztucznej inteligencji lub agenta i zintegrować go z ramami planowania przemysłu procesowego, aby zwiększyć zaufanie do decyzyjnych systemów AI poprzez poprawę ich przejrzystości i jakości decyzji? Metoda badawcza opiera się na Design Science (Hevner i in., 2004; Wieringa, 2014). Architektura referencyjna jest wspierana przez relewancję i rygor naukowy. Hevner et al. (2004) definiują siedem wytycznych dotyczących rozumienia, prowadzenia i oceny badań z zakresu design science. W aktualnych badaniach opracowano architekturę systemu referencyjnego dla godnej zaufania sztucznej inteligencji w planowaniu biznesowym w przemyśle przetwórczym (Galster, 2011; Nakaga-wa, 2014). Temat poruszany w tych badaniach ma ogromne znaczenie zarówno dla zastosowań teoretycznych, jak i praktycznych, o czym świadczą ostatnie badania. (Chiusi i in., 2020; Bejger & Elster, 2019; Bejger & Elster, 2020; Molnar, 2019; Singh i in., 2019; Tulio Ribeiro i in., 2016; Been Kim i in., 2016; Gunning, 2016; High-Level Expert Group on Artificial

Intelligence, 2019; Wachter i in., 2017; Chakraborti i in., 2020; Willms & Brandenburg, 2019; itp.) Ocena projektu została przeprowadzona w dwóch etapach. Po pierwsze, proces projektowania został oceniony w oparciu o zasady projektowania nakreślone przez Hevnera i in. (2004). Po drugie, przeprowadzono ankietę z ekspertami w dziedzinie architektury i planowania biznesowego w celu dalszej oceny projektu. Pytania ankietowe opierały się na kryteriach zaczerpniętych z najlepszych praktyk w badaniach i analizach. (Bass i in., 2021; Vasconcelos i in., 2005). Zostały one wprowadzone zgodnie z metodą opisaną przez Saunders i in. (2023) oraz Sekaran i Bougie (2019). Po przeprowadzeniu ankiety wyniki przeanalizowano przy użyciu metod statystycznych. Grupa ekspertów przedstawiła dodatkowe obawy, wymagania i ograniczenia, które zostały zidentyfikowane jako luki w tej pracy. Luki te zostaną uwzględnione w kolejnej iteracji cyklu projektowania. Wyniki ankiety potwierdzają hipotezę, że opracowana architektura referencyjna może zapewnić realne rozwiązanie określonego problemu. xi Powyższe Design-Science zostało wykorzystane do zaprojektowania i stworzenia architektury referencyjnej o nazwie "Re_fish" (połączenie słów "Rejewski" i "Babelfish", jako hołd dla Mariana Rejewskiego - czołowego polskiego naukowca, który złamał kod Enigmy, i Babelfish - fikcyjnego podmiotu i uniwersalnego dekodera dla każdej formy języka we wszechświecie), który może być wykorzystywany w kontekście planowania (korporacyjnego) w przemyśle przetwórczym, przy użyciu hybrydowego podejścia opartego na wiedzy (Hitzler i in. 2021; Hitzler & Sarker, 2022; Hochreiter, 2022; Niu et al. 2022; Futia & Vetrò, 2020; Marcus, 2020; Marcus & Davis, 2021; Sohrabi et al., 2018; Tiddi, et al., 2020).