

Nazwa w języku polskim: Wiarygodna Sztuczna Inteligencja
Nazwa w jęz. angielskim: Trustworthy Artificial Intelligence

Dane dotyczące zajęć:
Information on course:

Jednostka oferująca: Wydział Automatyki, Elektroniki i Informatyki // dr inż. Kamil Wereszczyński
Course offered by: Faculty of Automatic Control, Electronics and Computer Science// Dr. Kamil Wereszczyński

Język wykładowy:
polski
Language:
Polish
Strona WWW: Course homepage:
Skrócony opis:
Celem przedmiotu jest zapoznanie Studentów z uczeniem maszynowym i sztuczną inteligencją od strony prawno-etycznej i praktycznej. Kurs przeznaczony jest zarówno dla osób które są już zaznajomione z zagadnieniami uczenia się maszyn i sztucznej inteligencji, jak i dla takich, które takiej wiedzy nie mają. Studenci zapoznają się z: - podstawami uczenia maszynowego i sztucznej inteligencji – przypomnienie i rozszerzenie - zagadnieniami tzw. <i>wiarygodnej sztucznej inteligencji</i> (ang. <i>trustworthy AI</i>) w odniesieniu do zagadnień prawno-etycznych, - wiarygodnymi metodami trenowania i oceny wyników - wykorzystaniem istniejących metod (od „promptowania” po API) - implementacją SI w warunkach operacyjnych i zagrożeniami dla powodzenia projektów SI - budową zintegrowanych rozwiązań wykorzystujących sztuczną inteligencję i uczenie się maszyn
Short description:
The purpose of the course is to familiarize Students with machine learning and artificial intelligence from the legal, ethical and practical side. The course is intended both for those who are already familiar with the issues of machine learning and artificial intelligence, and for those who do not have such knowledge. Students will become familiar with: - basics of machine learning and artificial intelligence - reminder and extension - issues of so-called trustworthy artificial intelligence (trustworthy AI) in relation to legal and ethical issues, - trustworthy methods of training and results evaluation - use of existing methods (from “prompting” to APIs) - implementation of AI in operational settings and threats to the success of AI projects - building integrated solutions using artificial intelligence and machine learning
Opis:
Treści programowe: Wykład rozpocznie się od wprowadzenia do algorytmów uczących się, z naciskiem na podstawy uczenia maszynowego. Omówiona zostanie rola sztucznej inteligencji jako podzbioru uczenia maszynowego, w tym jej zastosowania oraz ograniczenia. Następna część wykładu skupi się na uczeniu głębokim, z uwzględnieniem jego rozwoju i wybranych architektur sieci głębokich, takich jak sieci konwolucyjne, transformery, sieci generatywne (GAN, GPT) oraz inne zaawansowane modele. Dodatkowo, zostanie omówiona fuzja danych jako istotny element w procesie tworzenia wiarygodnych modeli. Kolejny blok wykładu będzie poświęcony tematyce wiarygodnej sztucznej inteligencji. Zostanie wprowadzona definicja wiarygodności SI oraz omówione jej umocowania prawne i aspekty etyczne. Dyskusja obejmie także zagadnienia związane z ochroną praw autorskich i możliwe zagrożenia płynące ze stosowania sztucznej inteligencji, zarówno te oczywiste, jak i mniej intuicyjne. W dalszej części zostanie zaprezentowany proces ewaluacji metod sztucznej inteligencji, poczynwszy od etapów rozwoju pojedynczej metody SI, aż po omówienie przestrzeni ROC, czyli Receiver Operating Characteristic. W tej sekcji zostanie wyjaśniona definicja przestrzeni ROC oraz jej zastosowanie, a także

różne miary wynikające z tej przestrzeni. Omówione zostaną także przyczyny, dla których nawet doskonałe metody mogą zawodzić w praktycznych zastosowaniach.

Wykład obejmie również wiarygodną ewaluację metod sztucznej inteligencji, ze szczególnym uwzględnieniem procesu trenowania. Zostanie omówione, dlaczego jeden cykl trenowania nie wystarcza oraz jak walidacja krzyżowa może poprawić wiarygodność wyników. Przedstawione zostaną także błędy średniokwadratowe oraz inne miary, które pomagają zrozumieć skuteczność modelu w kontekście jego trenowania.

Kolejny etap wykładu poświęcony będzie finalnej ocenie wiarygodności algorytmów. Omówione zostaną metody sprawdzania, czy uzyskany wynik jest obiektywną oceną skuteczności algorytmu, czy też jest wynikiem przypadku. W tej części wprowadzone zostaną podstawowe metody statystyczne oraz reguła trzech sigm, służąca ocenie wyników algorytmów SI. Przykładem praktycznym będzie identyfikacja osobnicza na podstawie analizy wideo.

W dalszej części przedstawione zostaną metody wykorzystania gotowych rozwiązań sztucznej inteligencji. Zaprezentowany zostanie przegląd dostępnych serwisów SI, różnice między korzystaniem z „promptów” a API, a także ograniczenia metod dostępnych „z pudełka”.

Kolejny blok dotyczyć będzie kwestii budowy własnych rozwiązań SI. Omówione zostaną wady, zalety, koszty i czas związany z tworzeniem autorskich rozwiązań oraz wskazówki, kiedy i w jakim zakresie warto zdecydować się na budowę własnych systemów. Przedstawione zostaną także typowe etapy prac w projekcie AI oraz wymagania dotyczące zasobów sprzętowych i osobowych.

Następnie wykład poruszy temat akwizycji i adnotacji danych, w tym definicje związane z akwizycją, metody selekcji cech, rozszerzanie zbiorów danych oraz techniki adnotacji, zarówno automatyczne, jak i manualne. W tej sekcji zostanie omówione, jak korzystać z danych benchmarkowych i jak zapewnić ich odpowiednią jakość.

Na zakończenie wykładu zostanie przedstawiony sposób przygotowania i zaplanowania projektu sztucznej inteligencji, ilustrowany przykładem praktycznym z kancelarii podatkowej.

Wykład:

- **stacjonarne: 30 h**
- **Liczba punktów ECTS: 2**

Description:

Lecture:

The lecture will begin with an introduction to learning algorithms, with an emphasis on the basics of machine learning. The role of artificial intelligence as a subset of machine learning will be discussed, including its applications and limitations.

The next part of the lecture will focus on deep learning, including its development and selected deep network architectures, such as convolutional networks, transformers, generative networks (GAN, GPT) and other advanced models. In addition, data fusion will be discussed as an important element in the process of creating reliable models.

The next block of the lecture will be devoted to the topic of reliable artificial intelligence. The definition of trustworthy AI will be introduced and its legal entitlements and ethical aspects will be discussed. The discussion will also include copyright issues and possible risks of using artificial intelligence, both obvious and less intuitive.

In the next section, the process of evaluating artificial intelligence methods will be presented, starting with the stages of development of a single AI method, followed by a discussion of the ROC space, or Receiver Operating Characteristic. This section will explain the definition of the ROC space and its application, as well as the various measures derived from this space. The reasons why even excellent methods can fail in practical applications will also be discussed.

The lecture will also cover the reliable evaluation of artificial intelligence methods, with particular emphasis on the training process. It will be discussed why one cycle of training is not enough, and how cross-validation can improve the reliability of results. Mean squared errors and other measures that help understand the effectiveness of a model in the context of training will also be presented.

The next stage of the lecture will focus on the final evaluation of the reliability of algorithms. Methods for checking whether the result obtained is an objective assessment of the algorithm's effectiveness or the result of chance will be discussed. In this part, basic statistical methods will be introduced, as well as the three sigma rule for evaluating the performance of AI algorithms. A practical example will be personal identification based on video analysis.

In the next part, methods of using off-the-shelf artificial intelligence solutions will be presented. An overview of available AI services, the differences between using “prompts” and APIs, as well as the limitations of methods available “out of the box” will be presented.

The next block will address the issue of building your own AI solutions. The advantages, disadvantages,

costs and time involved in creating proprietary solutions will be discussed, as well as tips on when and to what extent it makes sense to decide to build your own systems. Typical stages of work in an AI project will also be presented, as well as requirements for hardware and personnel resources. Next, the lecture will touch on data acquisition and annotation, including definitions related to acquisition, feature selection methods, dataset expansion, and annotation techniques, both automatic and manual. This section will discuss how to use benchmark data and how to ensure its quality. The lecture will conclude with a presentation on how to prepare and plan an artificial intelligence project, illustrated with a practical example from a tax law firm.

Lecture:

- full-time studies: 30 h

Number of ECTS credits: 2

Literatura:

- B. Braunschweig and M. Ghallab, Eds., *Reflections on Artificial Intelligence for Humanity*, vol. 12600. in Lecture Notes in Computer Science, vol. 12600. Cham: Springer International Publishing, 2021. doi: [10.1007/978-3-030-69128-8](https://doi.org/10.1007/978-3-030-69128-8).
- H. Liu *et al.*, "Trustworthy AI: A Computational Perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, p. 4:1-4:59, Listopad 2022, doi: [10.1145/3546872](https://doi.org/10.1145/3546872).
- M. Ford, *Architects of Intelligence: The truth about AI from the people building it*. Packt Publishing, 2018. [Online]. Available: <https://books.google.pl/books?id=e4d7DwAAQBAJ>
- S. Lindgren, *Handbook of Critical Studies of Artificial Intelligence*. Edward Elgar Publishing, 2023. [Online]. Available: https://books.google.pl/books?id=5F_jEAAAQBAJ
- M. Trinh, *The AI project handbook: how to manage a successful artificial intelligence project*, First edition. in The artificial intelligence handbook series. New York [NY]: Rodeo Press, 2021. ISBN: 9798985117219 9798985117226

Bibliography:

- B. Braunschweig and M. Ghallab, Eds., *Reflections on Artificial Intelligence for Humanity*, vol. 12600. in Lecture Notes in Computer Science, vol. 12600. Cham: Springer International Publishing, 2021. doi: [10.1007/978-3-030-69128-8](https://doi.org/10.1007/978-3-030-69128-8).
- H. Liu *et al.*, "Trustworthy AI: A Computational Perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, p. 4:1-4:59, Listopad 2022, doi: [10.1145/3546872](https://doi.org/10.1145/3546872).
- M. Ford, *Architects of Intelligence: The truth about AI from the people building it*. Packt Publishing, 2018. [Online]. Available: <https://books.google.pl/books?id=e4d7DwAAQBAJ>
- S. Lindgren, *Handbook of Critical Studies of Artificial Intelligence*. Edward Elgar Publishing, 2023. [Online]. Available: https://books.google.pl/books?id=5F_jEAAAQBAJ
- M. Trinh, *The AI project handbook: how to manage a successful artificial intelligence project*, First edition. in The artificial intelligence handbook series. New York [NY]: Rodeo Press, 2021. ISBN: 9798985117219 9798985117226

Efekty uczenia się:

Wiedza: zna i rozumie podstawowe problemy współczesnej cywilizacji w odniesieniu do osiągnięć nauki i Techniki.

Umiejętności: potrafi samodzielnie planować i realizować własne uczenie się przez całe życie.

Kompetencje społeczne: jest gotów do krytycznej oceny posiadanej wiedzy i odbieranych treści, uznawania znaczenia wiedzy w rozwiązywaniu problemów poznawczych i praktycznych oraz zasięgania opinii ekspertów w przypadku trudności z samodzielnym rozwiązaniem problemu.

Learning outcomes:

Knowledge: knows and understands the basic problems of modern civilization in relation to the achievements of science and technology

Skills: is able to independently plan and implement his own lifelong learning

Social competence: is ready to critically evaluate the knowledge he possesses and the content he receives, to recognize the importance of knowledge in solving cognitive and practical problems, and to consult experts in case of difficulties in solving the problem independently.

Metody i kryteria oceniania:

Wykład:

Zaliczenie w formie uzyskiwania punktów aktywności.

Student może uzyskać punkty aktywności za

- Uczestnictwo w wykładach: 1 pkt za jeden wykład;
- Wykonywanie zadań domowych: 1-3 pkt za zadanie (max jedno per wykład)
- Prezentację wybranego zagadnienia na wykładzie: 1-15 pkt. (max jedno na cały kurs)
- napisanie eseju popularno-naukowego lub artykułu na wybrany temat: 1-15 pkt. (max jedno na cały kurs)

Zadania domowe są oceniane na zasadzie "wszystko albo nic": poprawny wynik oznacza przyznanie wszystkich punktów za zadanie, niepoprawny - zero.

Pozostałe są oceniane w skali: ndst. - 0 pkt., dst. - 7 pkt, +dst - 9 pkt, db -11 pkt, +db - 13 pkt, bdb -15 pkt.

Kryterium zaliczenia:

- od 8 pkt. dostateczny
- (8, 13) pkt. +dostateczny
- (13, 18) pkt. dobry
- (18, 23) pkt. + dobry
- od 23 pkt. bardzo dobry

Assessment methods and assessment criteria:

Lecture

Passing in the form of earning activity points.

Students can earn activity points for

- Attending lectures: 1 point per lecture;
- Completing homework assignments: 1-3 points per assignment (max one per lecture)
- Presentation of a selected topic in a lecture: 1-15 pts. (max one per entire course)
- Writing a popular science essay or article on a selected topic: 1-15 pts. (max one for the whole course)

Homework assignments are graded on an all-or-nothing basis: a correct score means the award of all points for the assignment, an incorrect score - zero.

The others are graded on a scale: 2.0 - 0 pts, 3.0 - 7 pts, 3.5 - 9 pts, 4.0 -11 pts, 4.5 - 13 pts, 5.0 -15 pts.

Passing criterion:

- from 8 pts. sufficient
- (8, 13) pts. +sufficient
- (13, 18) pts. good
- (18, 23) pts. +good
- from 23 pts. very good

Przynależność do grup przedmiotów w cyklach:

Element of course groups in various terms:

Opis grupy przedmiotów Course group description	Cykl pocz. First term	Cykl kon. Last term
przedmioty obieralne studia stacjonarne stopień studiów – dowolny kierunek studiów – dowolny, semestr dowolny elective courses full-time studies degree - any field of study - any semester - any	2024/2025	