UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

DEPARTEMENT D'INFORMATIQUE
DEPARTEMENT FÜR INFORMATIK
HUMAN-IST INSTITUTE
Bd de Pérolles 90
CH-1700 Fribourg

www.unifr.ch/inf

WORKING PAPER

Demystifying Artificial Intelligence

M. Mondal & D. Lalanne

Internal working paper no 21-02

October 2021

Demystifying Artificial Intelligence

Manuel Mondal
manuel.mondal@unifr.ch

Denis Lalanne
denis.lalanne@unifr.ch

Human-IST
University of Fribourg
October 2021

The objective of this document is to provide a non-technical introduction to the field of Artificial Intelligence. It is intended for anyone outside of the data science community curious about the subject, such as researchers from other areas looking for new instruments, legal professionals confronted with automated decision making algorithms or legal tech tools, entrepreneurs interested in expanding their products into new areas, etc. No prior knowledge of the domain is required for this introduction. External resources with technical details and additional examples are provided in the footnotes for the interested reader.

In the following sections, we introduce the fundamental concepts in the field, aim to clarify some commonly used terms and explain some limitations of modern AI techniques. First, we distinguish two types of AI: the symbolic approach, consisting of systems encoding rules and explicit knowledge, and the machine learning approach, which uses statistical methods and models to automatically infer relations between data and tasks to solve. In the subsequent section, we distinguish two complexity levels of these tasks. Finally, we introduce the subfield of "Explainable Artificial Intelligence" and distinguish three classes of explanations.

1 Types

Symbolic AI. Traditional AI systems are generally implemented as expert systems: a knowledge base of facts and a set of explicit logical relationships and formulas linking these facts. By applying the logical rules on the predefined facts from the knowledge base, new conclusions can be inferred[1].

This type of AI has mostly been abandoned at the turn of the century in favor of the statistical methods used in machine learning. Nonetheless, some usages persist[2] and there is an active research interest for hybrid systems combining the two approaches[3].

[1]For a more detailed introduction, consider Symbolic Reasoning (Symbolic AI) and Machine Learning.

[2]Consider the commercial example: Cyc.

[3]Consider (Garcez and Lamb 2020; Garnelo, Arulkumaran, et al. 2016; Garnelo and Shanahan 2019; Marcus 2020).

Machine Learning Systems. A plethora of end-consumers applications (translation tools, music recommendation engines, etc.) and media attention (successes and failures in autonomous driving technologies, DeepMind's victory against world champion Lee Sedol at Go, etc.) have brought modern machine learning algorithms into the limelight of public debate, economic interests, and regulatory oversight. Interest in this category of Artificial Intelligence systems has especially grown during the past two decades, even though its roots stretch back at least to the 1950s[4].

The various terms referring to specific techniques or subdomains in this field, such as *Deep Learning*, *Data Mining*, *Neural Networks*, *Pattern Recognition* and *Statistical Learning*, are often used interchangeably in the public discourse. For the purpose of this light introduction, we will not delve into the distinctions between these terms. We will use the umbrella term *Machine Learning Systems* when speaking of AI systems which rely on a fundamental principle of "learning by example" to solve tasks, i.e. which function by extracting salient patterns from training data, in order to tune the internal parameters of a statistical model. This definition of the term learning is notably formulated by (Mitchell et al. 1997):

> *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."*

The functioning of such systems can thus be summarized into three essential phases[5]:

(1) **Choosing a task to solve.** Common categories of tasks are *classification* (assigning a category to an input data point, e.g. recognizing objects in an image), *regression* (estimating a numerical value for an input data point, e.g. predicting a risk of recidivism), *natural language processing* (analyzing written texts, e.g. interpreting the meaning of a sentence) and *automated decision making* (e.g. automated credit decisions)[6].

Along with the task, a quality metric needs to be chosen, which will score the performance of the system (e.g. the number of misidentified objects in an image or the error on a numerical prediction).

The difficulty of the task determines the required size of the learning dataset (Figueroa et al. 2012; Halevy et al. 2009). This size can vary from a few hundred examples for simple problems to billions of data points for complex tasks such as autonomous driving or the training of sophisticated linguistic models[7].

(2) **The automated learning phase.** During this step, the examples are shown to the system in order to train the model. In the case of *Supervised Learning*, each example consists of

[4]For a more detailed technical introduction to the subject, consider (Raschka et al. 2020).

[5]Numerous steps were grouped here for simplicity. Consider process models such as CRISP-DM (Shearer 2000) for a detailed technical description of the relevant phases.

[6]For a more extended list of tasks consider (Goodfellow et al. 2016, section 5.1.1).

[7]A recent prominent example is GPT-3 (Brown et al. 2020), trained on 500 billion words, expressions, and phrases collected from the internet in order to learn how to hold a conversation.

a data point to evaluate (e.g. an image of a person's face) and the expected result (e.g. the name of the person.). These examples are processed by the system to automatically calculate the optimal parameters of a statistical model in order to (i) detect the important characteristics in each example and (ii) learn the relationship between these characteristics and the chosen objective.

In the case of *Unsupervised Learning*, it is not necessary to provide an expected result. The learning phase mainly consists in finding relevant statistical patterns in the dataset, such as groups of similar data points, anomalies, etc.

(3) **Model deployment.** Once the problem is modeled (phase 1) and the algorithm is trained with a sufficient number of examples (phase 2), the learned relationships allow the machine learning system to evaluate new examples presented without the expected result in order to predict it. Thus, during the training phase of e.g. a facial recognition AI, the system will have learned to detect the principal visual traits distinguishing individuals, based on the provided examples (phase 2). Once deployed, an image from a video surveillance system can be presented to this model (phase 3), which will try to identify the individual in the image using the traits learned in the previous phase.
Note that in a dynamic learning system, continuous learning and improvement is also possible, by iterating between the latter two phases.

In the case of structured data (e.g. personal profiles, criminal records) the characteristics taken into account to calculate the prediction can be manually selected before the learning phase in order to exclude some criteria (e.g. the ethnicity of a convict) during the *feature selection* phase[8]. However, the relative importance of the remaining features and their interaction effects during the final evaluation are determined automatically. Thus, intentionally excluding some criteria can be futile, especially if the same information can be inadvertently substituted by another property (e.g. the zip code unintentionally replacing the ethnicity in highly segregated areas[9]), or by implicitly deducing it from a combination of multiple other characteristics.

Even more so in the case of unstructured data (images, voice, natural language, etc.), where the characteristics detected by the machine will often differ from those intuitively considered by a person. Indeed, a facial recognition AI may not necessarily learn evident factors, such as eye color, the morphology of the faces, etc. but may consider more abstract properties, such as contrasts, color ratios imperceptible to the human eye or various other imprecise and unrecognizable artifacts[10].

The broader category grouping such phenomena is called *Shortcut Learning*, presented namely by (Geirhos et al. 2020). Much like a student only memorizing facts superficially to pass an exam with a minimal understanding of the subject, numerous deep learning systems suffer

[8]Feature selection and feature engineering are not only used in order to manually exclude unwanted criteria but are an essential step to setup any machine learning system.

[9]Consider e.g. (Zhang et al. 2017).

[10]Consider (Lin et al. 2020; Williford et al. 2020) for further investigations into the traits recognized by machines.

from an effect of shallow reasoning: optimizing for some unexpected artifacts, which allows the system to achieve good performances under very specific evaluation constraints. Consider one of their examples, in which a system developed to recognize pneumonia in radiographic images learned to spot specific hospital markers on the images instead, thus achieving good performance scores by identifying images from hospitals with high pneumonia prevalence rates[11].

In other words, the predictions and results of a machine learning system reflect the information it detects in the training data, without distinguishing between relevant signals and systematic biases[12]. Thus, decisions based on unexpected characteristics can at best be a tool for discovering previously unknown relationships. At worst, they can subtly and unwittingly reinforce the biases of a system[13]. All the aforementioned limitations relate to the topic of *Explainable AI*, introduced in the last section.

2 Tasks

The initial choice of the task to solve presupposes a distinction between types of tasks which can be solved by a machine, and those which cannot (yet).

Simple tasks. AI systems excel at solving tasks which (i) have clear, measurable, and explicit objectives, (ii) are constrained by known specifications and rules[14] and (iii) have a sufficiently large database of training examples available. Typical examples showcasing these criteria are games such as Go (Silver, Huang, et al. 2016), chess (McCarthy 1990) and video games (Badia et al. 2020). Indeed, for each of these examples (i) the objective can be quantified (a final score to maximize or intermediate probabilities to optimize), (ii) the set of possible worlds (the allowed player actions and potential situations) are limited and unchangeable and (iii) the system can learn from enough provided examples or can generate them by playing against itself (Silver, Hubert, et al. 2018).

Other usual AI tasks, such as image or facial recognition, voice recognition, automatic translation, rely on similar principles. For their functioning, enough training examples are ideally available to cover the whole spectrum of possible situations, their field of application is very specific and their performance can be objectively quantified.

Complex tasks. The usefulness of modern AI techniques is however exceeded when the considered task is not restricted to a domain with clearly defined limits. Thus, an autonomous driving system trained on usual traffic situations will be thoroughly confused by and unable

[11]Consider also the model presented by (Ribeiro et al. 2016), tasked to distinguish images of huskies from wolves, which learned to recognize the presence of snow in the background instead of learning the visual distinctions between the animals.

[12]Note that beyond biased datasets, most other stages of system design and deployment also raise ethical issues. Consider (Rochel and Evéquoz 2020) and (Elish and Boyd 2018) for detailed investigations of the subject.

[13]Consider the example of recidivism risk estimation: *Machine Bias* – ProPublica.

[14]Note however that some state-of-the-art systems are less and less dependent on having to know e.g. the rules of a game prior to learning how to play it (Schrittwieser et al. 2020).

to handle circumstances outside of the scope of the provided training data, such as a truck carrying traffic lights[15], uncommon billboards and sky-colored lorries crossing the highway.

Solving such tasks requires notably (i) knowledge of other domains (e.g. a facial recognition system will not be able to distinguish images of vehicles from buildings), (ii) a high-level understanding of the processed concepts (e.g. the same AI will not understand the notion of an individual or of a name either), (iii) common sense reasoning, (iv) intuitions, which were not, or cannot, be encoded, and (v) an understanding of the principle of causality. To use Judea Pearl's wording:

> *"All the impressive achievements of deep learning amount to just curve fitting".*[16]

With the progress in machine learning techniques, a growing number of tasks can be recategorized from the second type into the first. Nonetheless, overestimating one's system of the first type as one able to solve move complex tasks frequently causes issues with significant nefarious impact[17].

3 Intelligibility and explainability

A machine learning system is not always able to show its supervisor the reasoning behind a result or the justification for an automated decision: a phenomenon called the *black box* problem (Loyola-González 2019). Thus, in the example of recidivism risk prediction based on the personal profile of an inmate, many algorithms will not indicate which characteristics in his profile impacted the final estimated risk. However, to meet expectations of accountability, safety, liability, and responsibility, some level of inspection into the computation underlying a prediction must be made available (Danks and London 2017; Kingston 2016; Kroll et al. 2017). For this purpose (among others), the field of *Explainable AI (XAI)* has emerged and grown in recent years[18], providing transparency techniques for machine learning systems[19]. Various explanation techniques exist and can broadly be classified according to their level of access into the internal mechanisms of the AI system. We distinguish the methods below, broadly following the classification by (Guidotti et al. 2019) and (Du et al. 2019) for the latter three.

Formal explanations. As a first step, and without requiring additional technical efforts, the system designer can provide a formal description of the system's design and operation. For instance, he can indicate which learning algorithms were used to train the system, which dataset was used, how the dataset was examined for possible biases, and which automatic procedures

[15] Consider this video recording as an example: link.

[16] *To Build Truly Intelligent Machines, Teach Them Cause and Effect* – Quanta Magazine.

[17] Consider the long list of incidents collected in the AIAAIC repository – AI, Algorithmic and Automation Incident and Controversy Repository.

[18] For a systematic review of this domain, consider (Guidotti et al. 2019).

[19] The terms transparency, intelligibility, explainability, and interpretability are often used interchangeably by different authors and popular media. For details on this subject, consider (Clinciu and Hastie 2019).

were put in place to detect errors and problematic behavior once the model was deployed. This will provide the user or a system inspector the initial information needed to assess whether the system meets their expectations of rigor and reliability.

Post hoc explanations. This approach consists in reconstructing the reasoning of a machine learning system by analyzing its results. Examples:

— Which characteristics of a convict had an impact on the estimation of his recidivism risk (*local/outcome explanation*)?

— Are there population groups for whom the estimated recidivism risk is systematically higher than for other groups (*model explanation*)?

— Which would have been the risk of recidivism had the convict been older, or had belonged to another ethnic group (*counter factual explanation*)?

Such explanation techniques often require the construction of a second model in order to make the first one more intelligible, thus introducing a new issue of faithfulness between the initial AI system and the explanation model (Rudin 2019). The advantage of techniques in this category is the low level of access required to produce the explanations since providing examples of the system's inputs and outputs are generally sufficient.

Partial explanations. If beyond examples of the results, access to the internal mechanisms of the AI system is also available, more detailed explanations can be constructed, such as visualizations of intermediate results and processes, as well as more precise indications on the relative importance of the different criteria and their interaction effects (*model inspection*). A new risk of information overload is especially pronounced in this category: providing too fine-grained explanations may be distracting or even entirely misleading due to misunderstandings. The HCI (Human-computer interaction) research community is actively working on this issue (Alqaraawi et al. 2020; Wortman Vaughan and Wallach 2021).

Explainability by design. With this approach, the machine learning system is designed in such a way that the reasoning behind a prediction, as well as all the intermediate internal processes, are transparent and understandable by the supervisor, following the principle of *Transparent Box Design* (Rudin 2019). The construction of such a system, which is intrinsically interpretable, often requires a trade-off: by requiring additional efforts during its design and training or, sometimes, in the form of a loss in performance.

4 Conclusion

Artificial Intelligence is an ever-evolving domain, following a meandering path through the fields of logic, statistics, mathematics, and engineering. Its applications have the potential to change every field they touch, whether by deepening our understanding of the world, by automating procedures, or simply by complementing our own problem-solving abilities. The technological advances in computational power and the availability of large quantities of data during the past twenty years have brought the subfield of machine learning, and especially deep learning, to the attention of the wider public. In consequence, all fields adjacent to data analytics have become highly coveted in numerous industries. It is however unknown where future junctions in the road will take us, what will become possible over the next twenty years and which will be the dominant technical paradigm by that time. Yet undoubtedly, the field will continue impacting academic research, commercial interests, legislative efforts, and our daily lives for the foreseeable future. An understanding of the underlying concepts will thus be a necessity to stay on top of this evolution and will help to counteract the widespread trend of exaggerated claims in marketing campaigns and the co-opting of technical terms as meaningless buzzwords. With this document, we hope to have given the reader a first set of tools supporting them in this endeavor.

References

Alqaraawi, Ahmed, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze (2020). "Evaluating saliency map explanations for convolutional neural networks: a user study". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 275–285.

Badia, Adrià Puigdomènech, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell (Nov. 2020). "Agent57: Outperforming the Atari Human Benchmark". en. In: *International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, pp. 507–517.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell (2020). "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165*.

Clinciu, Miruna-Adriana and Helen Hastie (2019). "A Survey of Explainable AI Terminology". en. In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. Tokyo, Japan: Association for Computational Linguistics, pp. 8–13. DOI: 10.18653/v1/W19-8403.

Danks, David and Alex John London (Jan. 2017). "Regulating Autonomous Systems: Beyond Standards". In: *IEEE Intelligent Systems* 32.1. Conference Name: IEEE Intelligent Systems, pp. 88–91. ISSN: 1941-1294. DOI: 10.1109/MIS.2017.1.

Du, Mengnan, Ninghao Liu, and Xia Hu (May 2019). "Techniques for Interpretable Machine Learning". en. In: *arXiv:1808.00033 [cs, stat]*. arXiv: 1808.00033.

Elish, Madeleine Clare and Danah Boyd (2018). "Situating methods in the magic of Big Data and AI". In: *Communication monographs* 85.1. Publisher: Taylor & Francis, pp. 57–80.

Figueroa, Rosa L., Qing Zeng-Treitler, Sasikiran Kandula, and Long H. Ngo (Feb. 2012). "Predicting sample size required for classification performance". In: *BMC Medical Informatics and Decision Making* 12.1, p. 8. ISSN: 1472-6947. DOI: 10.1186/1472-6947-12-8.

Garcez, Artur d'Avila and Luis C. Lamb (Dec. 2020). "Neurosymbolic AI: The 3rd Wave". In: *arXiv e-prints* 2012, arXiv:2012.05876.

Garnelo, Marta, Kai Arulkumaran, and Murray Shanahan (2016). "Towards deep symbolic reinforcement learning". In: *arXiv preprint arXiv:1609.05518*.

Garnelo, Marta and Murray Shanahan (Oct. 2019). "Reconciling deep learning with symbolic artificial intelligence: representing objects and relations". en. In: *Current Opinion in Behavioral Sciences*. SI: 29: Artificial Intelligence (2019) 29, pp. 17–23. ISSN: 2352-1546. DOI: 10.1016/j.cobeha.2018.12.010.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann (Nov. 2020). "Shortcut learning in deep neural networks". en. In: *Nature Machine Intelligence* 2.11. Number: 11 Publisher: Nature Publishing Group, pp. 665–673. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00257-z.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. Vol. 1. 2. MIT press Cambridge.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi (Jan. 2019). "A Survey of Methods for Explaining Black Box Models". en. In: *ACM Computing Surveys* 51.5, pp. 1–42. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3236009.

Halevy, A., P. Norvig, and F. Pereira (Mar. 2009). "The Unreasonable Effectiveness of Data". In: *IEEE Intelligent Systems* 24.2. Conference Name: IEEE Intelligent Systems, pp. 8–12. ISSN: 1941-1294. DOI: 10.1109/MIS.2009.36.

Kingston, John (Dec. 2016). "Artificial Intelligence and Legal Liability". English. In: *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV*. Publisher: Springer-Verlag, pp. 269–279. DOI: 10.1007/978-3-319-47175-4_20.

Kroll, Joshua, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson, and Harlan Yu (Jan. 2017). "Accountable Algorithms". In: *University of Pennsylvania Law Review* 165.3, p. 633.

Lin, Yu-Sheng, Zhe-Yu Liu, Yu-An Chen, Yu-Siang Wang, Hsin-Ying Lee, Yi-Rong Chen, Ya-Liang Chang, and Winston H. Hsu (2020). "xCos: An Explainable Cosine Metric for Face Verification Task". In: *arXiv preprint arXiv:2003.05383*.

Loyola-González, Octavio (2019). "Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View". In: *IEEE Access* 7. Conference Name: IEEE Access, pp. 154096–154113. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2949286.

Marcus, Gary (Feb. 2020). "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence". In: *arXiv:2002.06177 [cs]*. arXiv: 2002.06177.

McCarthy, J. (1990). "Chess as the Drosophila of AI". en. In: *Computers, Chess, and Cognition*. Ed. by T. Anthony Marsland and Jonathan Schaeffer. New York, NY: Springer, pp. 227–237. ISBN: 978-1-4613-9080-0. DOI: 10.1007/978-1-4613-9080-0_14.

Mitchell, Tom M et al. (1997). "Machine learning". In.

Raschka, Sebastian, Joshua Patterson, and Corey Nolet (Apr. 2020). "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence". en. In: *Information* 11.4. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 193. DOI: 10.3390/info11040193.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (Aug. 2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

Rochel, Johan and Florian Evéquoz (Sept. 2020). "Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics". en. In: *AI & SOCIETY*. ISSN: 1435-5655. DOI: 10.1007/s00146-020-01069-w.

Rudin, Cynthia (May 2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". en. In: *Nature Machine Intelligence* 1.5. Number: 5 Publisher: Nature Publishing Group, pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

Schrittwieser, Julian et al. (Dec. 2020). "Mastering Atari, Go, chess and shogi by planning with a learned model". en. In: *Nature* 588.7839. Number: 7839 Publisher: Nature Publishing Group, pp. 604–609. ISSN: 1476-4687. DOI: 10.1038/s41586-020-03051-4.

Shearer, Colin (2000). "The CRISP-DM model: the new blueprint for data mining". In: *Journal of data warehousing* 5.4. Publisher: THE DATA WAREHOUSE INSTITUTE, pp. 13–22.

Silver, David, Aja Huang, et al. (Jan. 2016). "Mastering the game of Go with deep neural networks and tree search". en. In: *Nature* 529.7587. Number: 7587 Publisher: Nature Publishing Group, pp. 484–489. ISSN: 1476-4687. DOI: 10.1038/nature16961.

Silver, David, Thomas Hubert, et al. (Dec. 2018). "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". en. In: *Science* 362.6419. Publisher: American Association for the Advancement of Science Section: Report, pp. 1140–1144. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aar6404.

Williford, Jonathan R., Brandon B. May, and Jeffrey Byrne (2020). "Explainable Face Recognition". en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 248–263. ISBN: 978-3-030-58621-8. DOI: 10.1007/978-3-030-58621-8_15.

Wortman Vaughan, Jennifer and Hanna Wallach (Aug. 2021). "A Human-Centered Agenda for Intelligible Machine Learning".

Zhang, Lu, Yongkai Wu, and Xintao Wu (2017). "A Causal Framework for Discovering and Removing Direct and Indirect Discrimination". In: pp. 3929–3935.