Educational Administration: Theory and Practice
2023, 29(4), 2401 - 2408
ISSN: 2148-2403
https://kuey.net/

Research Article

An In-Depth Study Of Machine Learning In Artificial Intelligence

K. Jeevan Kumar[1*], K. Jairam[2], Ch. Ambedkar[3]

[1*]HOD – A.I & D.S. and A.I & M.L, V.K.R, V.N.B & A.G.K College of Engineering, Krishna District, Andhra Pradesh, India.
[2]Assistant Professor, S.R.K Institute of Technology, Enikepadu, Vijayawada, Krishna District, Andhra Pradesh, India.
[3]Assistant Professor, S.R.K Institute of Technology, Enikepadu, Vijayawada, Krishna District, Andhra Pradesh, India.

Citation: K. Jeevan Kumar, et.al (2023) An In-Depth Study Of Machine Learning In Artificial Intelligence, *Educational Administration: Theory and Practice*, 29*(4)* 2401 - 2408
Doi: 10.53555/kuey.v29i4.7127

ARTICLE INFO

ABSTRACT

Machine learning is a branch of artificial intelligence that enables algorithms to uncover hidden patterns within datasets, allowing them to make predictions on new, similar data without explicit programming for each task. Traditional machine learning combines data with statistical tools to predict outputs, yielding actionable insights. This technology finds applications in diverse fields such as image and speech recognition, natural language processing, recommendation systems, fraud detection, portfolio optimization, and automating tasks. For instance, recommender systems use historical data to personalize suggestions. Netflix, for example, employs collaborative and content-based filtering to recommend movies and TV shows based on user viewing history, ratings, and genre preferences. Reinforcement learning further enhances these systems by enabling agents to make decisions based on environmental feedback, continually refining recommendations. Machine learning's impact extends to autonomous vehicles, drones, and robots, enhancing their adaptability in dynamic environments. This approach marks a breakthrough where machines learn from data examples to generate accurate outcomes, closely intertwined with data mining and data science.

Introduction

In his US senate hearing in April 2018, MarkZuckerberg stressed the necessary capabilities of Facebook's "AI tools (...) to (...) identify hate speech(...)" or " (...) terrorist propaganda". Researchers would typically describe such tasks of identifying specific instances within social media platforms as *classification tasks* within the field of *(supervised) machine learning*. However, with risingpopularity of *artificial intelligence (AI)*, the term AI is often used interchangeably with machine learning–not only by Facebook's CEO in the exampleabove or in other interviews, but also across various theoretical and application-oriented contributions in recent literature. Carner (2017) even states that he still uses AI as a synonym for machine learning although knowing this is not correct. Such ambiguity, though, may lead to multiple imprecisions both in research and practice when conversing about methods, concepts, and results.

It seems surprising that despite of the frequent use of the terms, there is hardly any helpful scientific delineation. Thus, this paper aims to shed light on the relation of the two terms *machine learning* and *artificial intelligence*. We elaborate on the role of machine learning within instantiations of artificial intelligence, precisely within intelligent agents. To do so, we take a machine learning perspective on the capabilities of intelligent agents as well as the corresponding implementation.

The contribution of our paper is threefold. First, we expand the theoretical framework of Russel & Norvig (2015) by further detailing the "thinking" layer ofany intelligent agent by splitting it into separate "learning" and "executing" sublayers. Second, we show how this differentiation enables us to distinguish different contributions of machine learning for intelli gent agents. Third, we draw on the implementations of the execution and learning sublayers ("backend") to define a continuum between human involvement and agent autonomy.

In the remainder of this paper, we first review relevant literature in the fields of machine learning andartificial intelligence. Next, we present and elaborate our conceptual framework which highlights the contribution of machine learning to artificial intelligence.On that basis, we derive an agenda for future researchand conclude with summary, current limitations, as well as an outlook.

Related work

As a base for our conceptual work, we first review the different notions, concepts, or definitions of machine learning and artificial intelligence within extant research. In addition, we elaborate in greater detail on the theories which we draw upon in our framework.

Terminology

Machine learning and artificial intelligence, as well as the terms data mining, deep learning and statistical learning are related, often present in the same context and sometimes used interchangeably. While the terms are common in different communities, their particular usage and meaning varies widely.

\

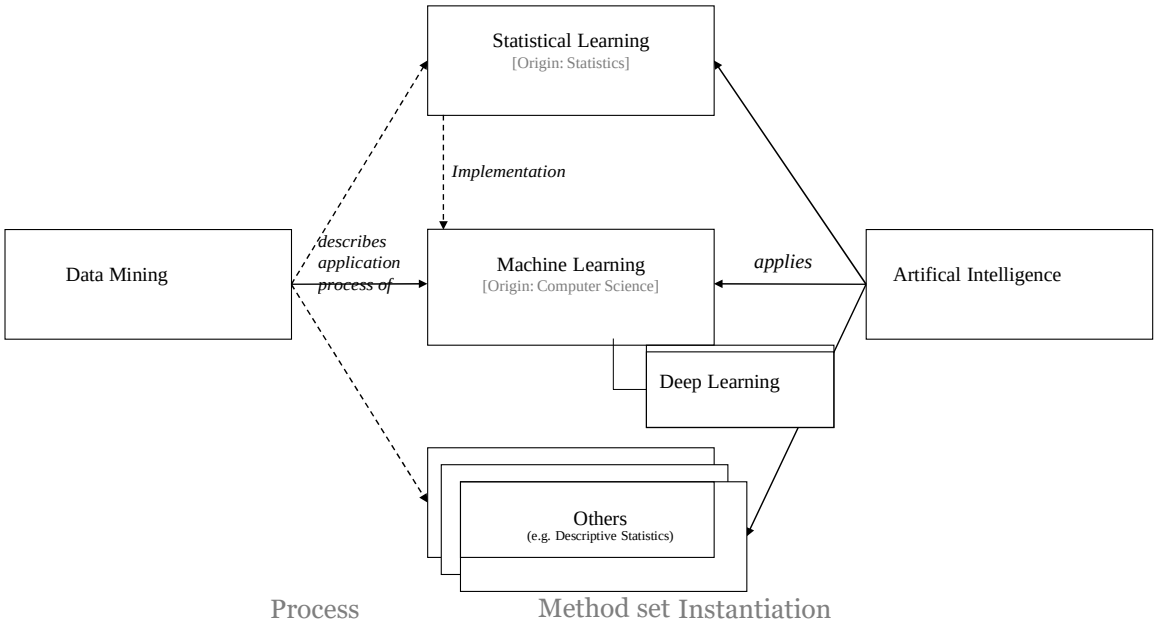

Figure 1. General terminology used in this paper

For instance, in the field of statistics the focus is on *statistical learning*, which is defined as a set of methods and algorithms to gain knowledge, predict outcomes, and make decisions by constructing models from a data set. From a statistics point of view, machine learning can be regarded as an implementation of statistical learning.

Within the field of computer science, *machine learning* has the focus of designing efficient algorithms to solve problems with computational resources. While machine learning utilizes approaches from statistics, it also includes methods which are not entirely based on previous work of statisticians—resulting in new and well-cited contributions to the field. Especially the method of deep learning raised increased interest within the past years. *Deep learning* models are composed of multiple processing layers which are capable of learning representations of data with multiple levels of abstraction. Deep learning has drastically improved the capabilities of machine learning, e.g. in speech or image recognition.

In demarcation to the previous terms, *data mining* describes the process on how to apply quantitative analytical methods, which help to solve real-world problems, e.g. in business settings. In the case of machine learning, data mining is the process of generating meaningful machine learning models. The goal is not to develop further knowledge about machine learning algorithms, but to apply them to data in order to gain insights. Machine learning can therefore be seen as a foundation for data mining.

Figure 1 and the terms defined within this paragraph lay the foundation of the remainder of this work. However, the overall terminology and relationships of the concepts is discussed controversially. Therefore, the focus of this paper is to bring more insight to the terminology and more precisely, to clarify the role of machine learning within AI. To gain a broader understanding for the terms machine learning and AI, we examine both in further detail.

Machine learning

Machine learning describes a set of techniques that are commonly used to solve a variety of real-world problems with the help of computer systems which can learn to solve a problem instead of being explicitly programmed. In general, we can differentiate between unsupervised and supervised machine learning. For the course of this work, we focus on the latter, as the most-widely used methods are of supervised nature. With regard to supervised machine learning, *learning* means that a series of examples ("past experience") is used to build knowledge about a given task. Although statistical methods are used during the learning process, a manual adjustment or programming of rules or strategies to solve a problem is not required. In more

detail, (supervised) machine learning techniques always aim to build a model by applying an algorithm on a set of known data points to gain insight on an unknown set of data. Thus, the processes of "creation" of a machine learning model slightly vary in their definition ofphases but typically employ the three main phases of model initiation, performance estimation and deployment: During the model initiation phase, a human user defines a problem, prepares and processesa data set and chooses a suitable machine learning algorithm for the given task. Then, during the performance estimation, various parameter permutations describing the algorithm are validated and a well-performing configuration is selected with respect to its performance in solving a specific task. Lastly, the model is deployed and put into practice to solve the task on unseen data.
Learning in general depicts a key facet of a human's cognition which "refers to all processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used" . Humans process a vast amount of information byutilizing abstract knowledge that helps us to better understand incoming input. Due to their adaptive nature, machine learning models are able to mimic thecognitive abilities of a human being in an isolated manner. However, machine learning solely represents a set of methods that enable to learn patterns in existing data, thus generating analytical models that can be utilized inside larger IT artifacts.

Artificial intelligence

The topic of artificial intelligence (AI) is rooted in different research disciplines, such as computer science, philosophy, or futures studies. In this work, we mainly focus on the field of computer science, as it is the most relevant onein identifying the contribution of machine learning to AI and in differentiating both terms.
AI research can be separated into different research streams. These streams differ on the one hand as to the objective of AI application (*thinking* vs. *acting*),on the other hand as to the kind of decision making (targeting a *human-like decision* vs. an *ideal, rational decision*). This distinction leads to four research currents which are depicted in Table 1.
According to the "Cognitive Modeling" (i.e. thinking humanly) stream, an AI must be a machine with a mind. This also includes performing human thinking, not only based on the same output as a human when given the same input, but also on the same reasoning steps which led to the very conclusion.
The "Laws of Thought" stream (i.e. thinking rationally) requires an AI to arrive at the rationaldecision despite what a human might answer.

Table 1. AI research streams based onRussell & Norvig

Objective Application to	Humanly	Rationally
Thinking	Cognitive Modeling	"Laws of thought "
Acting	Turing Test	Rational Agent

Therefore, an AI must follow the laws of thought by using computational models which reflect logic.
The "Turing Test" (i.e. acting humanly) stream implies that an AI must act intelligently when interacting with humans. To accomplish these tasks, an AI must perform human tasks at least as good as humans. These requirements can be tested by theTuring Test.
Finally, the "Rational Agent" stream considers an AI as a rational or intelligent agent. This agent does not only act autonomously but also with theobjective to achieve the rationally ideal outcome.
An alternative way to delineate AI is definingintelligence in general and using the resulting insights to create intelligent machines. Legg and Hutter use intelligence tests, theories of human intelligence and psychological definitions to define a measurement of intelligence. Based on their definition, they use an agent-environment framework to describe intelligence in general and—in case the agent is a machine—artificial intelligence in particular. Their framework exhibits many similarities to the "acting rationally" stream.
Besides defining AI in general, the classification of AI is another topic in the field of AI research. Searle suggests differentiating between weak and strongAI. Whereas a *weak AI* only pretends to think, a *strongAI* is a mind with mental states. Gubrud however categorizes AI by taking the type of task into account. An *artificial general intelligence* (AGI) is an AI which in general, i.e. in any domain, acts at least on the same level as a human brain, however without requiring consciousness. In contrast, a *narrow AI* is an AI that rivals or exceeds the human brain only in specific, limited tasks.
In the following, we will look into the "Rational Agent" stream in some more detail as it is of importance when regarding implementation of machine learning within AI. We will come back to theother three research streams in section 3 where we show that they are compatible with our framework of an agent-based AI.
According to the "Rational Agent" stream, the intelligence itself is manifested by the acting of agents. These agents are characterized by five features, namely they "operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals". An agent defines its action not for itself but with anenvironment it interacts with. It recognizes the environment by its sensors, has

an agent program to decide what to do with the input data, and performs an action with its actuators. To become a rational agent, the agent must also act to achieve the highest expected outcome according to this performance measure— based on the current and past knowledge of the environment and the possible actions.

When it comes to the general demarcation of agents, according to Russel & Norvig, the agent program can be segmented into four different agent types: A *simple reflex* agent reacts only based on its sensor data whereas a *model-based reflex* agent also considers an internal state of the agent. A *goal-based* agent decides for the best decision to achieve its goals. The fulfilment of a goal is a binary decision which means it can either be fulfilled or not. On the contrast, a *utility-based* agent has no binary goal but a whole utility function which it tries to maximize. An agent can become a *learning agent* by extending its program. Such a *learning agent* then consists of a performance element which selects an action based on the sensor data and a learning element, which gets feedback from the environment, generates own problems, and improves the performance element if possible.

The agent-environment framework consists of three components: an agent, an environment and a goal. Intelligence is the measurement for the "agent's ability to achieve goals in a wide range of environments". The agent gets input by perceptions generated from the environment. One type of perceptions are observations of the environment, while others are reward signals that indicate how well the goals of the agent are achieved. Based on these input signals, the agent decides to perform actions which are sent back as signals to the environment.

In order to understand the interplay of machine learning and AI, we base our concept on the framework of Russel & Norvig. With their differentiation between the two objectives of AI application, *acting* and *thinking*, they lay an important foundation.

Layers of agents

When trying to understand the role of machine learning within AI, we need to take a perspective which has a focus on the implementation of intelligent agents. We require this perspective, as it allows us to map the different tasks and components of machine learning to the capabilities of intelligent agents. If we regard the capabilities of *thinking* and *acting* of an intelligent agent and translate this into the terms of software design, we can reason that the *acting* capabilities can be regarded as a *frontend*, while the *thinking* part can be regarded as a *backend*. Software engineers typically strictly separate form and function to allow for more flexibility and independence as well as to enable parallel development. The frontend is the interface the environment interacts with. It can take many forms. In the case of intelligent agents it can be a very abstract, machine-readable web interface, a human-readable application or even a humanoid template with elaborated expression capabilities. For the frontend to interact with the environment, it requires two technical components; sensors and actuators. *Sensors* detect events or changes in the environment and forward the information via the frontend to the backend. For instance, they can read the temperature within an industrial production machine or read visuals of an interaction with a human. Actuators on the other hand are components that are responsible for moving and controlling a mechanism. While sensors just process information, actuators *act*, for instance by automatically buying stocks or changing the facial expressions of a humanoid. One could argue that the Turing test takes place at the interaction of the environment with the frontend, more precisely the combination of sensors and actuators if one wants to test the agent's AI of *acting humanly*. Despite every frontend having sensors and actuators, it is not of importance for our work what the precise frontend looks like; it is only relevant to note that a backend-independent, encapsulated frontend exists.

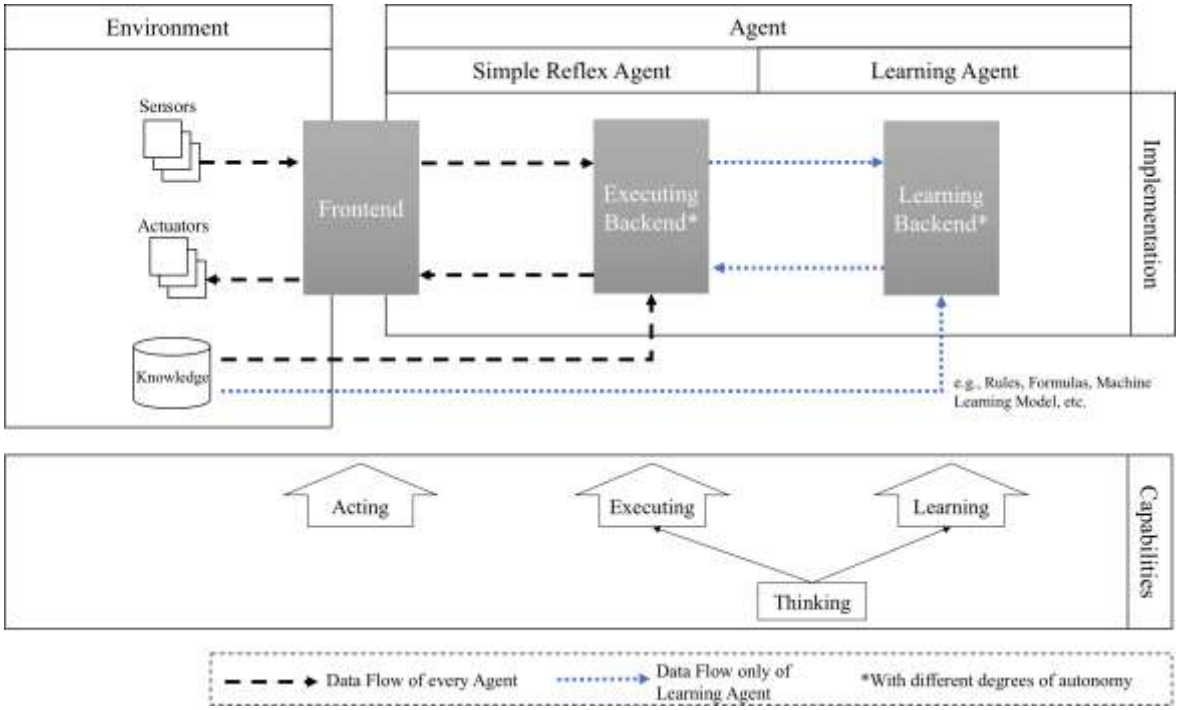

Figure 2. Conceptual framework

The backend provides the necessary functionalities, which depict the *thinking* capabilities of an intelligent agent. Therefore, the agent needs to learn and apply learned knowledge.
In consequence, machine learning is relevant in this implementation layer. When regarding the case of supervised machine learning, we need to further differentiate between the process task that is building (=training) adequate machine learning models and the process task that is executing the deployed models. Therefore, to further understand the roleof machine learning within intelligent agents, we refine the *thinking* layer of agents into a *learning* sublayer (model building) as well as an *executing* sublayer (model execution)[2]. Hence, we regard the necessary implementation for the learning sublayer asthe *learning backend*, while the executing sublayer is denoted by the *executing backend*.

Types of learning

The learning backend dictates first *if* the intelligentagent is able to learn, and, second, *how* the agent is able to learn, e.g., which precise algorithms it uses, what type of data processing is applied, how concept drift is handled, etc. Therefore, we pick up on theterminology from Russel & Norvig by regarding two different types of intelligent agents: *simple-reflexagents* as well as *learning agents*. This differentiationespecially holds for a machine learning perspective onAI, as it considers whether the underlying models in the *thinking* layer are once trained and never touched again (simple-reflex)—or continuously updated and adaptive (learning). In recent literature, suitableexamples for both can be found. As an example for simple-reflex agents, Oroszi and Ruhland build and deploy an early warning system of pneumonia in hospitals: While building and testing the model for the agent shows convincing results, the adaptive learning of the system after deployment might be critical. Other examples of agents with single-trained models are common in different areas, for instance for anaphora resolutions, prediction of pedestrians or object annotation. On the other hand, recent literature also gives examples for learning agents. Mitchell et al. present the concept of "never- ending learning" agents which have a strongfocus on continuously building and updating models within agents. An example for such an agent is shown by Liebman et al., who build a self-learning agent for music playlist recommendations. Other cases arefor instance the regulation of heat pump thermostats, an agent to acquire collective knowledge over different tasks or learning word meanings.
The choice on this feature in general (simple-reflexvs. learning agent) influences the overall design of theagent as well as the contribution of machine learning.The overview of our resulting framework is depicted in figure 2. In conclusion, in the case of a simple-reflexagent, machine learning takes places as a once-trainedmodel in the execution sublayer. In contrast, it plays arole in the learning sublayer of a learning agent to continuously improve the model in the execution sublayer. This improvement is based on knowledge and feedback, which is derived from the environmentvia the execution layer.

Continuum between human involvementand machine involvement

When it comes to the executing backend and the learning backend, it is not only of importance if and how underlying machine learning models are updated—but how much automated the necessary processes are. Every machine learning task involves various process steps, including data source selection, data collection,

preprocessing, model building, evaluating, deploying, executing and improving. While a discussion of the individual steps is beyond the scope of this paper, the autonomy and the automation of these tasks as an implementation within the agent is of particular interest in each necessary task of the machine learning lifecycle.

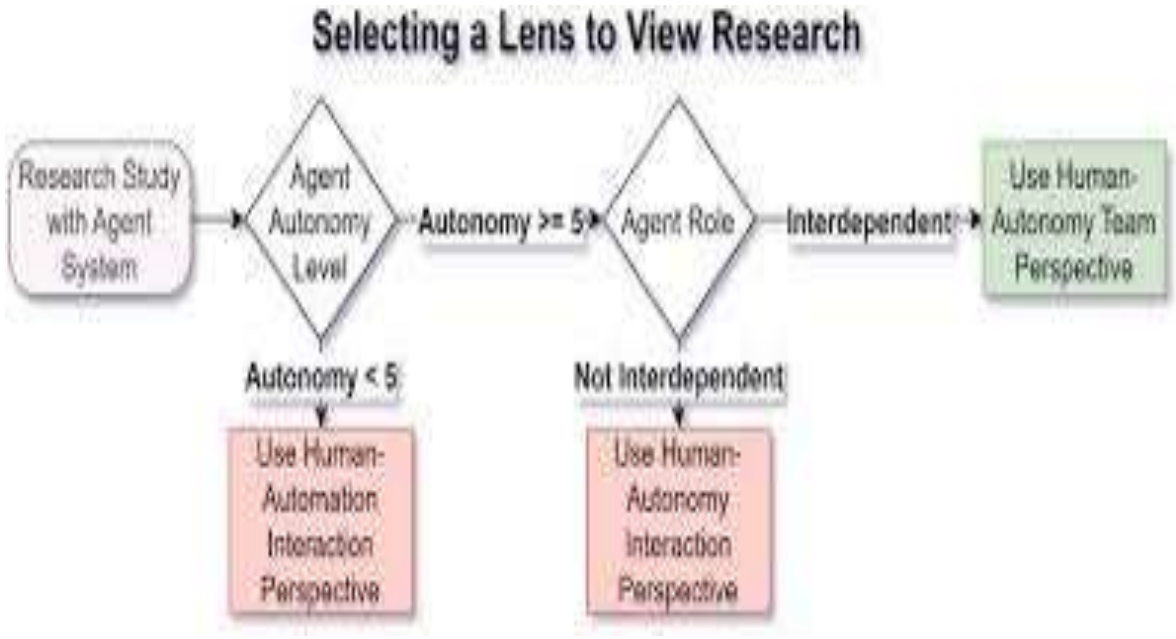

Figure 3. Degree of agent autonomy andhuman involvement

For instance, while the execution of a once-built model can be fairly easily automated, the automated identification of an adequate data source for a new problem or retraining as well as a self-induced model building are more difficult. Therefore, we need to view the human involvement in the necessary machine learning tasks of an intelligent agent, as depicted in figure 3. While it is hard to draw a clear line between all possible forms of human involvement in the machine learning-relevant tasks of an intelligent agent,we see this phenomenon rather as a continuum. The continuum ranges between none or little agentautonomy with full human involvement on the one extreme as well as the full agent autonomy and no or little human involvement for the delivered task on the other. For example, an intelligent agent with the task to autonomously drive a car considering the traffic signs already proves a high degree of agent autonomy. However, if the agent is confronted with a new traffic sign, the learning of this new circumstance might still need human involvement as the agent might not be able to "completely learn by itself". Therefore, the necessary involvement of humans, especially in the *thinking* layer (= executing backend and learning backend), is of major interest when describing AI andthe underlying machine learning models. The degree of autonomy for each step of machine learning can beinvestigated and may help to characterize the autonomy of an agent in terms of the related machine learning tasks.

Research priorities for machine-learning-enabled artificial intelligence

The presented framework of machine learning andits role within intelligent agents is still on a conceptuallevel. However, given the misunderstandings and ambiguity of the two terms [6–9], we see potential for further research with the aim both to clarify the terminology and to map uncharted territory for machine-learning enabled artificial intelligence.

First, empiric validation as well as continuous,iterative development of the framework is necessary. We need to identify various cases of intelligent agents across different disciplines and to evaluate how well the framework fits. It would be interesting to see howpractical and academic machine-learning-enabled artificial intelligence projects map to the framework, and, furthermore even quantify which share of such projects works with learning agents and which with non-learning agents. Additionally, such cases would help us to gain a better understanding of the necessary human involvement in state-of-the art intelligent agents—and, therefore, determine the "degree" of autonomy when regarding all aspects (acting, executing, learning) of such agents.

Second, one aspect of interest would be to reduce the necessary involvement of humans. As stated before, we see this spectrum as a continuum between human involvement and agent autonomy. Two possibilities come immediately to mind. The methods of *transfer machine learning* deal with possibilities on how to transfer knowledge (i.e., models) from one source environment to a target environment. This could indeed help to minimize human involvement, as further research in this field could show possibilities and application-oriented techniques to utilize transfer machine learning for automated adaption of novel or modified tasks.

Additionally, regarding already deployed models as part of the backend-layer, it is of interest not only how the models are built initially, but how to deal with changes in the environment. The so-called subfield of *concept drift* holds many possibilities on how to detect changes and adapt models—however, fields of successful application remain rare.

Conclusion

In a nutshell, machine learning models can be implemented as once-trained models within an intelligent agent—without the possibility to learn additional insights from the environment (simple reflex agent). Implementation-wise, we call this sublayer of executing knowledge the *executing backend*. In this case, the agent is able to utilize (previously built) machine learning models—but not build and update its own ones. If the agent, however, is able to learn from its environment and is, therefore, able to update the machine learning models within the execution sublayer, it is a learning agent. Learning agents have an additional sublayer, the *learning backend*, which allows them to utilize machine learning in terms of model building/training.

When it comes to the implementation of these two sublayers, it is of importance to capture the *degree of autonomy* that the machine learning within the agent requires. This aspect focusses on the human involvement in the necessary machine learning tasks, e.g. the data collection or the choice of an algorithm.

References

1. Aamodt, Agnar, and Enric Plaza. “Case-based reasoning: Foundational issues, methodological variations, and system approaches.” AI communications 7.1 (1994): 39-59.
2. Alain, Guillaume, and Yoshua Bengio. “Understanding intermediate layers using linear classifier probes.” arXiv preprint arXiv:1610.01644 (2016).
3. Alberto, Túlio C, Johannes V Lochter, and Tiago A Almeida. “Tubespam: comment spam filtering on YouTube.” In Machine Learning and Applications (Icmla), Ieee 14th International Conference on, 138–43. IEEE. (2015).
4. Ancona, Marco, et al. “Towards better understanding of gradient-based attribution methods for deep neural networks.” arXiv preprint arXiv:1711.06104 (2017).
5. Athalye, Anish, and Ilya Sutskever. “Synthesizing robust adversarial examples.” arXiv preprint arXiv:1707.07397 (2017).
6. Biggio, Battista, and Fabio Roli. “Wild Patterns: Ten years after the rise of adversarial machine learning.” Pattern Recognition 84 (2018): 317-331.
7. Breiman, Leo.“Random Forests.” Machine Learning 45 (1). Springer: 5-32 (2001).
8. Chen, Zhi, Yijie Bei, and Cynthia Rudin. “Concept whitening for interpretable image recognition.” Nature Machine Intelligence 2, no. 12 (2020): 772-782.
9. Cohen, William W. “Fast effective rule induction.” Machine Learning Proceedings (1995). 115-123.
10. Cook, R. Dennis. “Detection of influential observation in linear regression.” Technometrics 19.1 (1977): 15-18.
11. Deb, Kalyanmoy, Amrit Pratap, Sameer Agarwal and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” in IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, (2002).
12. Friedman, Jerome H, and Bogdan E Popescu. “Predictive learning via rule ensembles.” The Annals of Applied Statistics. JSTOR, 916–54. (2008).
13. Friedman, Jerome H. “Greedy function approximation: A gradient boosting machine.” Annals of statistics (2001): 1189-1232.
14. Ghorbani, Amirata, Abubakar Abid, and James Zou. “Interpretation of neural networks is fragile.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.
15. Ghorbani, Amirata, James Wexler, James Zou and Been Kim. “Towards automatic concept-based explanations.” Advances in Neural Information Processing Systems 32 (2019).
16. Grömping, Ulrike. “Model-Agnostic Effects Plots for Interpreting Machine Learning Models.” Reports in Mathematics, Physics and Chemistry: Department II, Beuth University of Applied Sciences Berlin. Report 1/2020 (2020)
17. Heider, Fritz, and Marianne Simmel. “An experimental study of apparent behavior.” The American Journal of Psychology 57 (2). JSTOR: 243–59. (1944).
18. Holte, Robert C. “Very simple classification rules perform well on most commonly used datasets.” Machine learning 11.1 (1993): 63-90.
19. Inglis, Alan, Andrew Parnell, and Catherine Hurley. “Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models.” arXiv preprint arXiv:2108.04310 (2021).
20. Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. “Feature relevance quantification in explainable AI: A causal problem.” International Conference on Artificial Intelligence and Statistics. PMLR (2020).

21. Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. "Concept bottleneck models." In International Conference on Machine Learning, pp. 5338-5348. PMLR (2020).
22. Laugel, Thibault, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. "Inverse classification for comparison-based interpretability in machine learning." arXiv preprint arXiv:1712.08443 (2017).
23. Linsley, Drew, et al. "What are the visual features underlying human versus machine vision?." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017.
24. Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490, (2016).
25. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017).
26. Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. "Anchors: high-precision model-agnostic explanations". AAAI Conference on Artificial Intelligence (AAAI), 2018
27. Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. (2020).
28. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. "Definitions, methods, and applications in interpretable machine learning." Proceedings of the National Academy of Sciences, 116(44), 22071-22080. (2019).
29. Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks." Advances in neural information processing systems 29 (2016): 3387-3395.
30. Nguyen, Anh, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. "Plug & play generative networks: Conditional iterative generation of images in latent space." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4467-4477. 2017.
31. Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ACM (2017).
32. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." AAAI Conference on Artificial Intelligence (2018).
33. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." ICML Workshop on Human Interpretability in Machine Learning. (2016).
34. Robnik-Sikonja, Marko, and Marko Bohanec. "Perturbation-based explanations of prediction models." Human and Machine Learning. Springer, Cham. 159-175. (2018).
35. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. (2017).
36. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).
37. Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.