Ralf Möller

Humanities-Centred Artificial Intelligence (CHAI) as an Emerging Paradigm

Abstract: This essay argues that there needs to be a shift in artificial intelligence, namely, a shift from the perspective of computation-oriented machine learning to the perspective of application-oriented machine training. With a training perspective on artificial intelligence rather than a learning perspective, issues raised by the humanities will become ever more important in the development of new artificial intelligence systems; from a training perspective, the goal is that intelligent systems act in ethical ways, solve tasks in trustworthy ways, and that system configuration be controlled by humanities researchers during a basic training phase. The central idea of Humanities-Centred AI (CHAI) is that intelligent systems, proven to be a useful support to humanities researchers in solving their humanities-research problems, will also be suitable for other application domains because the models used internally by these systems contain sufficient information about human culture. Humanities researchers can even actively shape these models, which can then be shared among agents and reused in different contexts. A prerequisite is that artificial intelligence moves away from offering so-called artificial intelligence methods for computer scientists and instead provide useful abstractions in the form of intelligent agents whose instruction or training is controlled by humanities researchers.

1 Introduction

Artificial intelligence (AI) is the science that studies the synthesis and analysis of intelligent systems as part of application systems and services. In AI terminology, intelligent systems are composed of agents.¹ Before we analyse what intelligent systems actually are, we must first examine this notion of an intelligent agent. Agents can be mere software systems or, e.g. humanoid robots equipped with respective hardware. From a technical point of view, the behaviour of human beings interacting with intelligent systems can, in principle, also

1 Russell and Norvig 2020; Poole and Mackworth 2017; Shoham and Leyton-Brown 2012.

be abstracted as agent behaviour. As we will see in this article, the standard notion of computational correctness² applies to simple programs but neither to humans nor to computational intelligent agents.

To benefit from AI, humans communicate goals to agents, which, roughly speaking, then interpret and adopt goal specifications as precisely as possible. Then, agents fulfil respective goals as accurately as possible with a sequence of dedicated actions that agents compute by means of, for instance, planning algorithms,³ while continuously checking whether the underlying goals for the planning process are still based on valid assumptions. To fulfil their goals, agents try to automatically acquire suitable models. While many AI proponents tend to say that agents acquire ‘knowledge’ and that they ‘learn’, basically what is accomplished is that models are built on the basis of sequences of inputs (‘percept’ sequences) that the agent receives or acquires from its host environment. Note that models used by an agent can be declarative, i.e. models can have a formal semantics and can be queried in various ways, or the models can be merely computational, i.e. they can be directly executed with specific parameter settings such that computational results are obtained which are required for fulfilling goals on the one hand and for interacting with humans in an anticipatable way on the other. Algorithms for answering queries with respect to a model often become ‘inference’ or ‘reasoning’ algorithms such that one can avoid mathematical formality. The old satirical saying, ‘Do not anthropomorphise computers; they really hate that’, is certainly applicable in this context. In order to keep the text of this article non-technical, we also use respective notions for discussing artificial intelligence in general and agents in particular, but we emphasize that quite simple, non-magical technical definitions of terms such as ‘inference’ or ‘reasoning’ can be given with reference to formal mathematical models.

Interestingly, when speaking about intelligent agents, it is indeed rather hard to avoid the terms ‘knowledge’ or ‘learning’. Learning as a subfield of AI has developed tremendously in recent years due to advances in computer technology⁴ as well as in mathematics.⁵ While early learning approaches still relied on vast amounts of labelled training data (pairs of input and expected output for supervised learning), today even unsupervised learning is of enormous practical relevance, e.g. in building language models for processing text data in

² Dijkstra 1968; Hoare 1969.

³ LaValle 2006; Ghallab, Nau and Traverso 2016.

⁴ Soyata 2020.

⁵ Deisenroth, Faisal and Ong 2020.

various languages. Relevant features of data are no longer hand-crafted beforehand but are determined automatically during learning (learning of representations). Even one-shot or few-shot learning has become an aspirational research topic lately.⁶

Rather than focusing on single agents, we introduce the notion of an intelligent system. An intelligent system is a system that can be composed of one agent or a set of agents. In the latter case, a system is comprised of a mechanism that sets up the rules for agent cooperation. No matter what kind of intelligent system is used, it has become clear that the performance of intelligent systems of the future will heavily rely on well-selected data for building and operating a system. In the very near future, we will see a switch from technical issues of machine learning in computer science as they are prominently discussed now to the more general topic of systematic machine training, with a focus on the selection, preparation, and curation of appropriate teaching material for intelligent systems (and single agents) such that certain non-trivial tasks can be carried out successfully. Humanities research, e.g. in the area of philology, will become an important driving factor for AI applications that use natural-language processing technology, in which agent reasoning is language based.⁷

Example

Let us consider an information retrieval example in which we imagine an agent that can find relevant documents for helping a scholar solve a humanities research problem. The agent must be instructed about the problem-related field of expertise of the scholar. [...] The agent plans a sequence of actions (i.e. its own actions and the possible actions of the scholar) that provide the means for posing follow-up queries or for providing reinforcement feedback. The utility measure of the agent is assumed to be maximised if the scholar is satisfied with the (final) retrieval result.

We will see an interesting shift from today's engineering-based AI to a future Humanities-Centred AI because the notion of machine training will be dominant in the future, and new research areas in the humanities will emerge that deal with how to advise agents, e.g. information retrieval agents. Scholars will select training material (e.g. reference libraries) for agents and specify web-mining goals to prepare agents for specific tasks.

⁶ Yao 2021.

⁷ MacCartney 2009.

Example

For instance, a researcher might want to obtain scientific publications and data for a research project concerning cuneiform clay tablets but limit the publications to those that are already summarized and appropriately linked to relevant X-ray or CT data of clay tablets.

The vision is that a humanities scholar will educate a web-mining agent so that it can accomplish this task successfully. While building a model for the scholar's information-retrieval problem, the agent internally constructs so-called subjective content descriptions ('semantic annotations') of the documents in the reference library. The descriptions (or annotations) are subjective because the agent tries to describe both the reference library's content (e.g. with symbolic descriptions of entities identified in the text) and the model of the scholar's research problem provided by the reference library (i.e. for other scholars, different content descriptions might be generated). The reference library can be updated by the humanities scholar, and she or he can check whether the agent produces better answers than for previous (and hopefully future) information-retrieval queries. The better the reference library gets (and the more focused the string queries are), the better the agent is instructed about the nature of the research problem to be solved.

Well-organised machine-readable descriptions of, e.g. written artefacts and other humanities research topics will allow AI systems to reduce their present brittleness in communication with humans by anticipating at least some aspects of human thinking and behaviour.

Example

The (symbolic) subjective content descriptions provide a (rough) representation of the knowledge of a certain domain from a subjective point of view.⁸ Those subjective content descriptions that help solve research problems (indicated to the agent with reinforcement feedback) can then be used to generate a formal model of a problem domain via machine-learning techniques. The agent can then exploit this model to support other humanities scholars in the future. Providing a reference library for instructing an agent in a particular task therefore trains the agent in solving other problems as well. The better the reference libraries are composed, and the more that relevant problems are solved in co-operation with the agent, the better the agent is trained, and the better the services of the agent will become.

⁸ Cf. Moretti 2016.

Only if training material is carefully selected by humanities researchers dealing with language and cultural artefacts in an appropriate way, will agents become trustworthy service providers. Agents that are trained by humanities researchers in the anticipated way, i.e. with the right literature at the right time, will also be enormously helpful for systematically constructing, e.g. chatbots to be used in application contexts that are completely different compared to the initial context of finding relevant documents for solving a humanities research problem as envisioned at the beginning of this example.

In the following paragraphs, we analyse some requirements for making this vision of Humanities-Centred AI a reality. In order to draw up a roadmap for investigating this topic, we need to delve a little bit deeper into AI as a field of research and its public perception.

2 On the self-conception of contemporary artificial intelligence

The essential assumption of AI is that agents are situated in an environment and that by perceiving this environment, agents can construct (or learn) a model of it (and possibly of other agents in that same environment). Agents act *rationaly* by calculating optimal actions (or sequences of actions) from their local point of view, based on their current goals.⁹ Computed actions, when executed, may exert an influence on the environment, the state of which must then be captured again by the agent so that it can continue to act optimally (cyclic interaction with the environment). Goals can be hierarchically structured and prioritised.

Example

Let us consider an agent playing chess (i.e. a chess program). The agent analyses the chess board, and, when it's the agent's turn, computes the best move by recursively speculating what the opponent could do, what the best move would be then, and so on. Note that because the search space is so immense that the agent can hardly find an optimal move for a complex mid-game board position, the agent needs to cut the search short and make estimations about its winning chances in a particular situation. Once a

⁹ Eisenführ, Weber and Langer 2010.

move is selected, the agent executes the move (or just announces the move to the environment, which then executes the move). Chess is a simple agent scenario because every agent has perfect information about the state of the environment (the board), and all moves have deterministic effects. The assumption of perfect information being available, however, is not necessarily useful in most settings.

The acquisition of information about the environment can also include feedback regarding the choice of past actions by the agent, and this feedback may cause the agent to reassess its internal state and its planning for future action (reinforcement).¹⁰ Sequential-decision theory solves the problem of determining the best action even when the state as well as the effect of actions are uncertain,¹¹ but theory assumes that both the distribution of events in the environment and all possible actions are known a priori, which is usually an inappropriate assumption in real-world settings. Over time an agent can estimate, e.g. the distribution of certain representative values for describing the environment, and the agent also has to conceive new actions on the fly (e.g. during a web-mining mission). When planning optimal actions, every agent is subject to the laws of complexity theory, i.e. the rationality of agents is quite limited, and not only under strict time constraints. In other words: optimal actions cannot necessarily be determined with available computational resources (time and memory). Since optimal actions can sometimes only be calculated approximately, it has to be accepted that the theoretically achievable goals of an agent are not always achieved or are only achieved in a delayed fashion.

The described form of system construction with situated agents, which from their local point of view act optimally in an environment, i.e. act rationally, forms the basis for an artificial concept of intelligence. An agent is considered intelligent if, through situated interaction with the environment, it achieves its goals in an unexpectedly short time (from the perspective of a human observer that interprets the agent's behaviour from the outside). Some formalisations of AI even require that goals be achieved in rather different contexts (or environments), based on abstract descriptions (a set of rules) that an environment provides to an agent, e.g. rules of different games such as Chess, Go, or Shigo. For each of these games, it has been shown that with reinforcement learning¹² a system can be indeed generated just from a description of the rules of the game

¹⁰ Sutton and Barto 2018.

¹¹ Hastie and Dawer 2010.

¹² Sutton and Barto 2018.

(artificial general intelligence or general game playing).¹³ Even more impressive is that generated systems can easily beat human players. It should be emphasised, however, that games are definitely very specific problems. The generated systems will almost always win against human players, but they cannot explain to a beginner how to become a good player, which clearly shows the limits of contemporary AI. Nevertheless from the analysis of solutions to very specific problems, general techniques for the analysis and synthesis of intelligent systems can indeed be derived.

We emphasize that the attribution of intelligence depends on the expectations of the human being who interprets the behaviour of computational agents.

Example

If, for example, a chess agent (or chess program) has the goal of winning 1000 games, and achieves this goal in a little more than 1000 games, 1010 games, say, and one does not expect it to achieve this goal after 1010 games, then one might call this chess program intelligent.

An information-retrieval agent that provides decent information-retrieval results with only few interaction cycles, when given a reference library and some form of query, would also be considered intelligent.

A pocket calculator, however, is expected to correctly calculate 1000 out of 1000 addition operations; therefore the calculator program is hardly ever called intelligent.

If people expect a program to always win or to always compute the correct result, then they tend not to call that program intelligent. Needless to say, system-performance expectations can certainly change over time.

2.1 Misunderstandings

In any case, the concept of intelligence is poorly defined without taking into account the situation in which an agent is working. It is also important to understand that the intelligence of agents only shows itself from the outside. That is, the intelligence of agents, which we humans read into their behaviour as software or hardware systems, is achieved with classical techniques of

¹³ Genesereth and Thielscher 2015.

computer science. Classifying a computational method as an AI method does not really lead to deep insights. There are no intelligent algorithms or intelligent data. There are, however, intelligently developed algorithms and intelligently collected (or composed) data. But the latter tasks are accomplished by human experts. Although we often read it in the press, it does not make much sense to speak about the generation of an AI or the integration of an AI into a system. Using the term ‘Intellectics’¹⁴ would perhaps have been more appropriate for the scientific field of artificial intelligence.¹⁵

It is also important to understand that in AI the term ‘learning’ primarily refers to an agent building a model of its environment *at runtime*: i.e. the agent applies learning algorithms to data from its ‘sensors’ to better achieve its goals (besides exploiting reinforcing feedback from the environment). The use of learning techniques at the developmental stage of agents to generate an initial model or behaviour is nevertheless certainly possible, but using learning techniques in this way does not hit the core of AI and can also be understood as data science for the development of agent systems. Knowledge-based learning or model-based learning can be used at development time as well as at runtime.¹⁶ Learning by using prior knowledge is indeed essential to reducing necessary training data, because for the vast majority of learning problems it is quite difficult to provide sufficient training data at runtime – a fact that is often discussed, but hardly really tackled today. With prior knowledge, it becomes possible to automatically structure the hypothesis space for possible learning outcomes, which is a necessary prerequisite for effective agents. Speaking in terms of ‘educating an agent,’ we see that training includes taking care of at runtime (not only at setup time) so that the agents learn the right thing. We could, then, extend the notion of machine training to ongoing machine training,¹⁷ and the most appropriate sequence of training steps, to be defined by humanities researchers, definitely matters in this context.

A common misconception is to assume that in AI, learning (at runtime) is only relevant in relation to the models used (e.g. for improving functionalities such as the classification of images, recognition of objects, etc.). It is equally important that agents analyse their input data (perceptual data) in order to develop certain strategies to deal efficiently with the very limited resources,

¹⁴ <<http://www.intellektik.de>> (accessed on 30 June 2021).

¹⁵ Hölldobler 2000.

¹⁶ Bishop 2013.

¹⁷ Continual learning also means that existing functionality which has proven to be beneficial is not overwritten by new learning activities.

with respect to time and space, for determining the best next action. With techniques such as automated machine learning (AutoML),¹⁸ i.e. automatic algorithm configuration and parameterisation, very efficient strategies for processing sequences of percepts can be automatically generated or learned after a short training phase. There are application scenarios in which agents first process simple instances of a problem class (e.g. solving satisfiability [SAT] problems for Boolean formulas) in order to identify suitable strategies for solving harder problems from the respective problem class. Only after warming up can difficult instances of a problem class be tackled successfully.

Example

For instance, solving Sudoku problems might be reduced to solving the satisfiability of Boolean formulas (automatically derived to encode the original Sudoku problem). The encoding formulas have a certain structure, and initially solving easier Sudoku problems can help the solver adapt to the specific formulas and develop dedicated strategies to solve the satisfiability problem for Sudoku-specific Boolean formulas, such that more difficult Sudoku problems (resulting in more difficult SAT problems) can also be solved.

Without adaptation through simple problems, it may be that difficult problems cannot be processed at all with given time constraints without running into a timeout. Learning to perform inference strategies can be much more important in practical applications than learning functionality. This kind of learning scenario is also part of the training metaphor; one must be able to evaluate (using challenge tasks) whether an agent is ready for the next steps involving more complex tasks.

2.2 Applications

Agents are not explicitly modelled in the software in every application. In some applications, only parts of the cycle described above – perceiving, concluding, acting – can be recognised. It turns out that the AI perspective manifested in the agent metaphor has also produced generalizable principles that are useful for standard software development (e.g. learning a classifier for photogrammetric images).

¹⁸ Hutter, Kotthoff and Vanschoren 2018.

Algorithms used to program systems by means of learning techniques are very popular. Nevertheless in current applications mostly only a part of the basic idea of artificial intelligence is explicitly reflected.

Example

For example, techniques of signal-level speech processing (perception) are useful for developing chatbots, which may have hardly any application knowledge defined in an explicit form and may hardly determine actions in an optimal way as described above, but which still provide useful services in web applications.

Audiological signal processing can be used in robots as well as in technical solutions for new hearing aids.

Techniques for the evaluation of, for instance, photogrammetric image data (perception) together with application knowledge implicitly integrated into the software make it possible to highlight critical areas in the image so that a new application can be set up to provide optimised support in making decisions (e.g. about whether an area is polluted or not).

It may be that only parts of the perceiving-concluding-acting cycle are manifested in the image-processing software or hardware. Nevertheless one can still speak of an AI application, i.e. an application of data-processing techniques developed in AI. This kind of scenario is an example of good old-fashioned AI, since technicians and computer science experts are still required in this context, a situation that we should try to overcome in order to make AI more useful to, e.g. humanities scholars.

2.3 Fears about AI?

Note that, unlike humans, intelligent artificial agents that have achieved high performance by interacting in a specific environment can be copied and can continue to learn and improve themselves in a different environment. Increasing the competence of agents by learning through interaction with an environment is a very interesting principle of application design. Learning at the time of system development is usually not enough. Whether a system created by means of learning techniques can be operated in a certain environment without continuous learning is a question that cannot be answered in general. Indeed, reality

is evolving, and without continuous adaptation the usability of systems created by means of learning techniques is certainly limited.

Fears about the results of the science of AI could very well originate from overestimating the efficiency of the aforementioned chain-letter-like development of the intelligence of agents, especially if one considers the inherently limited rationality of agents. The point at which systems emerge that are more intelligent than humans (however such intelligence could be measured) is sometimes called the Singularity.¹⁹ The creation of intelligence, especially through the automatic combination of self-learning subsystems, is not a reality today, in particular because the rationality of agents is necessarily limited due to complexity considerations. The spontaneous use of a newly registered contact in an address book by a speech-based assistant, i.e. a functionality that already exists today, is not very impressive from this perspective. The correct pronunciation of names is also no witchcraft, due to the large amounts of speech data that are available today. Continuous learning must be organised, however, and it must be part of the planning for agent training. Thus, with the notion of machine learning changed to machine training, there is much more control about what kind of capabilities an intelligent agent will be able to achieve, such that misgivings of the kind introduced above can be addressed.

It must be made clear, however, that mass influence on people through intelligent agents is already a reality today, and without countermeasures, will become increasingly subtle so that democracies may even be put at risk. In addition, one must not confuse intelligence with human-compatible behaviour,²⁰ because the (unexpectedly) fast achievement of goals does not mean that the goals being set up independently by an agent are reasonable from a human perspective. If human control over setting targets in automatic systems is lost, the future might turn into a nightmare. The science of AI must concentrate its research on not only enabling the creation of systems that are correct in a certain sense and controllable, but also on creating systems that are provably human-compatible. It is rightly argued that agents must take into account not their own goals but the goals or preferences of people so that a positive benefit can be demonstrably achieved for the people involved in agent interaction. Initially uncertain about people's preferences and the goals supplied by humans, agents reduce this uncertainty through situated interaction. An agent should maintain permanent uncertainty about human preferences and goals in order to anticipate past mistakes and to enable adaptation to changing human preferences.

¹⁹ Kurzweil 2005.

²⁰ Russell 2019.

Expressions such as ‘human-allied AI’, ‘human-aware AI’, ‘human-compatible AI’ have been coined in order to capture these ideas.²¹

2.4 The ethics of AI as a field of science

Ethical questions related to research in the humanities play an important role in AI concerning the training perspective. For example, instead of understanding ethics simply by giving designers of systems moral guidelines formulated in natural language, AI offers the possibility of using formal ethics to enable intelligent agents to model moral action so as to estimate and weigh people’s preferences. This idea represents an enormous gain in comparison with ethical approaches that are too short-sighted and applied only for system construction. The idea opens up important perspectives.²² It is not only the designers of AI systems who must act according to ethical guidelines but also the constructed systems (agents) themselves,²³ and continuously educating those systems should systematically ensure moral action. Ethical behaviour of intelligent agents will only emerge if, when developing the mechanism of a system (environment and interaction of agents), it is taken into account that acting according to ethical guidelines is a profitable (or even dominant) strategy for each individual agent. To design such a mechanism is not easy, and agent training also matters. The AI research area ‘algorithmic mechanism design’²⁴ investigates the basics of this design challenge. The area must be extended to take continuous machine training into account. Note that the approach of following ethical guidelines when designing systems is not wrong, but such an approach does not fully meet the needs arising from the potential of AI. Rational actions of agents must also comply with certain ethical guidelines. Other aspects relate to fairness, privacy, and security:²⁵ for example, the ability to determine whether or not particular data (such as the data of a particular person) was part of the training data, or at least was the basis for answering an enquiry, ought to be avoided. Techniques such as differential privacy²⁶ are highly relevant to trustworthy agent systems²⁷ and should be technically combined with ethical

²¹ See also Kambhampati 2020.

²² Cf. terms such as ‘value alignment’ (Russell 2019).

²³ Russell 2019.

²⁴ Nisan and Ronen 2001; Nisan et al. 2007.

²⁵ Kearns and Roth 2020.

²⁶ Li et al. 2016.

²⁷ Marcus and Davis 2019.

considerations. The consideration of privacy sets new research goals for artificial intelligence as a science.

An agent's objectives (goals) may be uncertain, and since objectives are reflected in the tasks assigned to an agent, its tasks exhibit uncertainties from the agent's viewpoint, too (and also from the human's viewpoint if the human is conceptualized as an agent). Thus, there is hardly any standard notion of correctness that is applicable to single agents. What can be analysed, however, is a system of agents interacting in some environment. Analysing a system of agents (i.e. a so-called mechanism) and examining whether a desired functionality is actually realised by such a system of agents is a central issue in the science of artificial intelligence. Rather than asking for formal correctness, the main question is whether a mechanism maximises social welfare understood as a (combined) measurement of global utility for all agents involved (or at least whether the mechanism provides a decent value for social welfare). A mechanism that maximises social welfare is called appropriate. Without a determination as to the appropriateness of a mechanism in a concrete case, intelligent systems simply might not be applicable in certain areas. It seems that this ethical issue of AI is sometimes neglected in practical applications. Humanities researchers should contribute ethical competence to discussions about appropriateness.

2.5 Perspectives

Research in the field of AI aims to identify the methods and techniques of computer science that are necessary for developing and analysing verifiably human-compatible intelligent agents. In order to produce intelligent agents, computer scientists have developed novel programming and modelling techniques such as probabilistic differential programming as a generalised view of deep learning. Probabilistic-relational fundamentals of modelling real-world systems, techniques of data analysis (data science, data mining), and optimising decision making on the basis of large amounts of data (possibly with web scaling) are just as fundamental for system creation in AI. New approaches such as imputative learning are intended to, among other things, solve problems related to acquiring models (= learning) from *incomplete* data. The results of these efforts can later be incorporated into, e.g. in-depth language comprehension, which is important for boosting the learning capabilities of agents. Computer scientists, however, can only partially contribute to agent learning, namely, with respect to the technical basis. What is required is a way to let humanities researchers act as machine trainers. This perspective constitutes a roadmap for a new view of AI.

3 Humanities-centred AI: A roadmap

While technology always evolves, it is important that, on the one hand, Humanities-Centred AI (CHAI for short) concentrates on an appropriate training interface such that AI with machine-training capabilities becomes effective for the humanities. On the other hand, computer scientists must adopt the right perspective from which to design trainable systems. In the following paragraphs, we will first analyse some traps that AI researchers have run into, and then describe the CHAI roadmap.

Let us assume that someone implements a system that can do something that humans can also do, e.g. play a shooter game. Further, for the implementation of the technical system, among other things, method X is applied. Let us then further assume that the resulting shooter-game system is quite close to human performance, maybe even better. A human being, here a human player, is assumed to be intelligent by definition, and therefore the technical shooter-game system is also called intelligent. Underlying this conclusion is a very old definition of AI: a system is intelligent if it does something for which a human (apparently) needs intelligence. Unfortunately, this definition has many pitfalls and is no longer suitable for defining the subject matter of AI as a science today.

3.1 First trap

In the above definition, method X becomes a so-called AI method if method X can be successfully used in an application context in which humans solve problems (e.g. playing shooter games, playing chess, as well as in other application contexts) in order to generate systems that can simulate, support, or replace humans. The temptation to talk about AI methods is particularly strong when one thinks that the technology behind a method is, in whatever vague way, analogous to the physiological processing of stimuli in humans, or that the technology can be seen as analogous to the human level of information processing (a level assumed to be high), or even that the technical behaviour is in some way reminiscent of human problem-solving behaviour. An example of the first analogy is the application of so-called neural networks, which, however, as a technology have become extremely distant from the physiological original over time and are actually purely mathematical in nature. But some people also use the expression ‘AI method’ for other seemingly cognitively analogous methods, such as the application of rule-chaining systems to the creation of so-called expert systems. Some people even think that the expression ‘AI method’ refers

to arbitrary architectures for solving problems for which humans need intelligence. This loose use of the expression ‘AI method’ leads to the second trap.

3.2 Second trap

If a developer applies method X, which she or he calls an AI method, to develop a system for application problem P, which many humans have never solved, cannot solve, or can solve only after a great deal of very specific training, then the developer often draws the erroneous conclusion that she or he is working on an AI problem, simply because she or he is applying a so-called AI method. The trap in this case is that a problem is classified by the solution method.

Example

An example is automatically parking automobiles without stopping other traffic. This is a problem that very few people can solve. Because one can solve the problem (in a rather poor way) with dedicated technical sensor devices, for which we do not find analogues in human beings, a researcher might claim to have solved an AI problem, but even if one of the so-called AI methods is employed, parking is still far from being an AI problem. A problem simply cannot be classified according to the solution method.

Let us consider another example. The analysis of multispectral images is an interesting application problem. The problem can be approached with methods that are also suitable for the development of intelligent game systems (such as Asteroids), i.e. with supposed AI methods, here perhaps the training of so-called convolutional networks (rather deeply cascaded, linear-filter operations, combined with respective nonlinear aggregations). But taking a method that is useful for an intelligent system and applying that method to solving a technical, image-interpretation problem does not mean that image interpretation is an AI problem; it is an (interesting and difficult) application problem.

3.3 Third trap

The system that is realized via the alleged AI method X, e.g. a function parameterized by techniques of algorithmic learning (i.e. a convolutional network in actual, final use, with all its concretely determined parameter values), is called artificially intelligent by some people. Oddly enough, some people even call the

system ‘an artificial intelligence’, or assert that ‘an AI’ has been created. Unfortunately, this way of thinking would require us to call every function intelligent (or even ‘an artificial intelligence’) if it was created with method X, called an AI method, which is absurd.

Traps of this kind can only be mitigated by not defining AI in human terms, but by focusing on the nature of the problems being solved.

A specific AI problem is actually to find algorithms for evaluating a percept and for determining actions so that the goals are achieved as completely and as quickly as possible over a long period of time, with the result that the agent can act rationally. Machine training, however, should be accomplished not by computer scientists but by domain experts. It should be noted that the laws of computer science apply to any agent (and to humans) no matter how training is organised: the computation time is not in every case sufficient for computing optimal actions (bounded rationality). If the agent, however, achieves its goals with unexpected speed, despite these intrinsic limitations, it is said to be intelligent. Now, a real AI problem is to enable domain experts to construct or generate such intelligent agents, without computer scientists or dedicated experts in machine learning being in the loop (the computer scientists or dedicated experts then deal with good old-fashioned, technical AI methods).

These kinds of real AI problems are studied in CHAI and are solved by whatever method. It is not the methods, techniques, or algorithms that distinguish AI in this context, but the problems solved, namely, real AI problems. The same consideration applies to other sciences: it is not problem-solving methods that define the science of physics, e.g. methods for solving differential equations, but rather the problems being solved (or phenomena studied).

Example

For example, the aforementioned shooter game, seen as an agent, has the goal of winning 1000 out of 1000 games. The actions selected by the program (i.e. where to manoeuvre the spaceship depending on the situation, or whether to shoot) will contribute to achieving the goal.

If the agent achieves the goal frequently, it will, from the outside, be considered intelligent, unless it is expected to achieve the goal all the time. However, if the agent internally uses a function parameterized by evaluating a large set of input-output examples (i.e. an algorithmically learned function) that the developer generated, then it is not the function that is intelligent. The entire agent can be considered intelligent from the outside only if it exhibits the appropriate behaviour using the function. A simple set of learned, nested functions that

implement a convolutional network, for example, cannot be intelligent because the functions do not compute actions and do not interact with an environment at all, but are only (externally) applied and then return results. Nevertheless the functions can certainly make an agent that uses them behave in an intelligent way, according to the definition given above. The third trap is one that people fall into who do not distinguish between an internally used function and an agent. When people talk about creating ‘an AI,’ they usually (unconsciously) mean creating an intelligent agent. Unfortunately, in such cases no agent is generated, but only a simple function that cannot be intelligent; this situation again reflects the absurdity of speaking about generating ‘an AI’.

Thus we see that the developer of an agent, not the agent itself, solves an AI problem. The agent solves an application problem – unless we build agents that construct agents. The former agents then also solve an AI problem. But we are not yet that far. Even so, by emphasizing the notion of machine training by domain experts via CHAI, we can avoid falling into the aforementioned traps.

Example

Although an agent that analyses multispectral images can be intelligent (it does its job quickly and really well), we have argued that it solves not an AI problem, but an application problem. It is not the analysis of a multispectral image that is an AI problem, but the development of the agent that can analyse these image data (correctly and quickly), by, e.g. quickly achieving its goal of correctly classifying 1000 out of 1000 multispectral images.

Image classification is just one example. Another task in which the agent can help is finding multispectral images for a specific research question (recommendation task). Another task could be to combine multiple images and spectra so that an added value is achieved. Training agents to support researchers in such tasks is out of reach these days, but would be very powerful. The knowledge required for such tasks could be gathered by first having the agent solve information-retrieval problems in the humanities.

The method for building the agent, e.g. a method for training a convolutional network, is therefore not an AI method, but a method for solving application problems. If a researcher operates as a computer scientist to train the network (and specifies the network architecture) but does not develop an agent, then the researcher in question is not working in AI, but in applied computer science. Working as a computer scientist in solving application problems is of course a worthwhile activity, but we should not refer to that activity as solving AI problems. Computer scientists working in CHAI develop AI technology that

allows domain experts (humanities scholars) to be involved in machine training for problem solving.

3.4 New developments: Abstractions rather than methods

Besides machine training, AI must develop techniques so that users can inform agents in an appropriate way about goals (or tasks), or even so that the agents can automatically determine goals in an appropriate way. Task solutions, in turn, must be communicated to humans by agents in an appropriate way. What does it mean to communicate in a way that is appropriate for humans? It means that AI supporting the machine-training perspective (and not only the machine-learning perspective) must also explore aspects of human information processing in order to meet the demands. Here again, a fundamental difference with statistics or the field of data science becomes clear: it is not only ongoing learning and (optimal) actions of agents that are added to the overall picture, but also the anticipation of human information processing and thus also of human knowledge. If these issues are irrelevant in a particular application context, then it is better to pragmatically speak of data evaluation, statistics, or – with an industrial twist – data science. From this perspective, the results of research in the humanities, e.g. in linguistics, philosophy, or cultural studies, are very relevant for AI.

Developing an agent, however, may very well take place via certain methods that have been or are being developed in the science of AI. These methods include the following: methods for correctly setting up learning machinery (e.g. ensuring independent and identically distributed data as preconditions of learning); methods for verifying the properties of models used by agents (e.g. discriminatory power determination); or methods for proving the appropriateness of an agent's actions in any situation (e.g. addressing ethical concerns). Such methods might then usefully be called AI methods. With AI methods, one solves AI problems (see above), not application problems.

Several questions arise:

- Should humanities scholars solve AI problems?
- Should humanities scholars work in applied computer science?

The answer to both questions is no. Humanities scholars are trained to solve humanities problems, not applied computer-science problems or even specific AI problems. If humanities scholars become applied computer scientists, the chances increase that they either cease to be good humanities scholars, or they

will acquire the necessary computer-science skills very late in their career and thus could hope to work effectively with their own graduates or even colleagues who do not have these skills. So overall effectiveness would suffer.

Now, what is the charm of AI anyway? What are the hopes associated with it? One hopes that humanities scholars will not have to solve applied computer-science problems in order to solve a humanities problem but will be able to define, with reasonable effort, agents' goals that will then solve the humanities application problems. The important question that this aspiration raises is:

- For which humanities problems can basic agents be provided, so that humanities researchers can define (or refine) the agents' goals with reasonable effort and the agents can solve the problem for the humanities researcher or in interaction with the humanities researcher?

The right question for designing an AI training scenario in the humanities is therefore:

- What would a humanities researcher need to know in order to be able to select agents, train them, and then appropriately define the goals for the selected agents?

Some relevant questions for AI research from the CHAI perspective are:

- How can agents be designed so that they can be trained by humanities researchers, such that the training improves solving a certain problem and progress can be properly evaluated by the trainer?
- Can AI system developers enable humanities researchers to define agent goals with reasonable effort and thus to solve humanities application problems themselves without having to become applied computer scientists?
- How can (pre-)trained models be found and exchanged?
- How can models that have been constructed for a certain task by training be used successfully for related tasks?

The goal of collaboration between AI scientists and humanities scholars should not only be the immediate solution of humanities application problems. The interaction of humanities scholars and AI scholars should also generate new agent types so that the goals of the agents that belong to the newly developed type are easier for humanities scholars to define and so that application problems in the humanities can be solved in a truly sustainable way (by humanities scholars). It is clear that humanities application problems will be used as examples throughout this work on new agent types. Nevertheless solving these application problems alone should not be the primary goal, because if it were, then

only short-term success is likely to be achieved. Involving natural scientists adds even more complexity to the overall picture.

AI as a science has yet to demonstrate that it can provide trainable agent types whose goals can be set by humanities researchers with reasonable effort, so that humanities application problems can really be solved by humanities scholars. The same deficit is even truer for collaboration with natural scientists. Learning-technology, assumed to be universally applicable, could help to specify functions used internally in an agent. But if an agent is to act and communicate with a human about achieving certain goals, this is just the beginning. Other techniques for specifying or aligning goals must be provided in a complementary fashion, keeping in mind that goals are almost always underspecified and must be specified as appropriate in the interaction. But if goals are to be communicated, then they can hardly remain implicit in the agent's design, as is still almost always the case in simple systems used today (see the comments on the agent for multispectral-image analysis).

3.5 Adopting a new perspective: From machine learning to machine training

As discussed at the beginning of this essay, there is potential for a shift from the technical issues of machine learning to the humanities issues of machine training (i.e. educating a machine), such that intelligent systems can solve tasks relevant for humanities research, are trainable by domain experts, solve tasks in a trustworthy way, and act in an ethical way. For all cases, humanities research provides the scientific background. A prerequisite is that we move away from focusing on so-called AI methods for computer scientists to providing abstractions in the form of intelligent agents and respective mechanisms for coordinating the agents, while supporting the view of machine training rather than machine learning.

Example

In our example concerning information retrieval, problem-specific documents will be carefully collected and then provided to an agent as a reference library so that domain-specific expertise is provided to the agent in a useful way in a preparatory training context. The agent can then try to find additional resources, either from its previous services or from other agents on the web. Thereafter, the agent can search for relevant publications on the web that match the query of a human. Based on feedback in the interaction cycle, which will have been anticipated by the agent during planning,

the agent can reason about its strategies for inferring the problem-specific information needed by the human researcher, and the agent can evaluate and adapt its provision of documents in the reference library, including the addition of subjective content descriptions to be used for information retrieval in that context.

Whether AI is already powerful enough to provide something useful in this sense is the ultimate question. I conjecture that AI scientists will have to keep working hard. But they at least understand the true AI problem that they need to solve, namely, supporting machine training by well-educated humanities scholars or natural scientists rather than supporting a technical focus on machine learning controlled by computer scientists only. If humanities as a field is going to exploit agent technology, and I have argued that it should, then humanities scholars ought to be involved in CHAI as early as possible so that they can derive a deep understanding of what the agent can actually achieve on the one hand, and provide important insights for AI researchers on the other; in the CHAI approach advocated here, these insights will concern a training perspective rather than a learning perspective.

Acknowledgements

I am indebted to Jost Gippert for insightful comments on a draft of this essay. In addition, for discussions that helped shape the ideas behind this contribution, I would like to thank the team of the DFG Research Unit ‘Data Linking’, which is part of the Cluster of Excellence ‘Understanding Written Artefacts’ at University of Hamburg, namely, Marcel Gehrke, Sylvia Melzer, Stefan Thiemann, and Simon Schiff, as well as Tanya Braun from the Institute of Information Systems at University of Lübeck.

References

- Bishop, Christopher M. (2013), ‘Model-based machine learning’, *Philosophical Transactions A (Mathematical, Physical and Engineering Sciences)*, 371/1984: <doi.org/10.1098/rsta.2012.0222>.
- Deisenroth, Marc Peter, A. Aldo Faisal and Cheng Soon Ong (2020), *Mathematics for Machine Learning*, Cambridge: Cambridge University Press.
- Dijkstra, Edsger (1968), ‘A constructive approach to the problem of program correctness’, *BIT Numerical Mathematics*, 8/3: 174–186.
- Eisenführ, Franz, Martin Weber and Thomas Langer (2010), *Rational Decision Making*, Berlin: Springer.

- Genesereth, Michael and Michael Thielscher (2015), *General Game Playing* (Synthesis Lectures on Artificial Intelligence and Machine Learning, 24), San Rafael, CA: Morgan & Claypool.
- Ghallab, Malik, Dana Nau and Paolo Traverso (2016), *Automated Planning and Acting*, Cambridge: Cambridge University Press.
- Hastie, Reid and Robyn M. Dawer (2010), *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*, 2nd edn, Los Angeles, CA: SAGE.
- Hoare, C. Anthony R. (1969), 'An Axiomatic Basis for Computer Programming', *Communications of the ACM*, 12/10: 576–580.
- Hölldobler, Steffen (ed.) (2000), *Intellectics and Computational Logic: Papers in Honor of Wolfgang Bibel* (Applied Logic Series, 19), Dordrecht: Springer.
- Hutter, Frank, Lars Kotthoff and Joaquin Vanschoren (eds) (2018), *Automated Machine Learning: Methods, Systems, Challenges*, Cham: Springer.
- Kambhampati, Subbarao (2020), 'Challenges of Human-Aware AI Systems', *AI Magazine*, 41/3: 3–17.
- Kearns, Michael and Aron Roth (2020), *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*, Oxford: Oxford University Press.
- Kurzweil, Ray (2005), *The Singularity is Near: When Humans Transcend Biology*, New York, NY: Penguin.
- LaValle, Stephen M. (2006), *Planning Algorithms*, Cambridge: Cambridge University Press.
- Li, Ninghui, Min Lyu, Su Dong and Weining Yang (2016), *Differential Privacy: From Theory to Practice* (Synthesis Lectures on Information Security, Privacy, and Trust, 18), San Rafael, CA: Morgan & Claypool Publishers.
- MacCartney, Bill (2009), *Natural Language Inference*, PhD dissertation, Stanford University.
- Marcus, Gary and Ernest Davis (2019), *Rebooting AI: Building Artificial Intelligence We Can Trust*, New York, NY: Vintage Books.
- Moretti, Franco (2016), *Distant Reading*, Konstanz: Konstanz University Press.
- Nisan, Noam and Amir Ronen (2001), 'Algorithmic Mechanism Design', *Games and Economic Behavior*, 35/1–2: 166–196.
- Nisan, Noam, Tim Roughgarden, Éva Tardos and Vijay V. Vazirani (2007), *Algorithmic Game Theory*, Cambridge: Cambridge University Press.
- Poole, David and Alan Mackworth (2017), *Artificial Intelligence: Foundations of Computational Agents*, 2nd edn, Cambridge: Cambridge University Press.
- Russell, Stuart (2019), *Human Compatible: AI and the Problem of Control*, New York, NY: Viking.
- Russell, Stuart and Peter Norvig (2020), *Artificial Intelligence: A Modern Approach*, 4th edn, Boston, MA: Pearson.
- Shoham, Yoav and Kevin Leyton-Brown (2012), *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*, Cambridge: Cambridge University Press.
- Soyata, Tolga (2020), *GPU Parallel Program Development Using CUDA*, s.l.: Chapman & Hall.
- Sutton, Richard S. and Andrew G. Barto (2018), *Reinforcement Learning: An Introduction*, 2nd ed, Cambridge, MA: MIT Press.
- Yao, Fupin (2021), *Machine learning with limited data*, <arXiv:2101.11461> (accessed on 30 June 2021).