dai

Proceedings of the First Conference on
Designing with Artificial Intelligence
dai digital 2020

Proceedings of the First Conference on
Designing with Artificial Intelligence
Online conference, September 17-19, 2020
https://www.designing-artificial-intelligence.eu

Edited by
Marc Engenhart
Sebastian Loewe

Published by
appliedAI, an initiative for the application of AI technology
(www.appliedai.de)
Digital Product School
(www.digitalproductschool.io)

Both are part of
UnternehmerTUM
Lichtenbergstraße 6
Munich-Garching
D-85748, Germany

Identifier
ISBN 978-3-00-068634-4

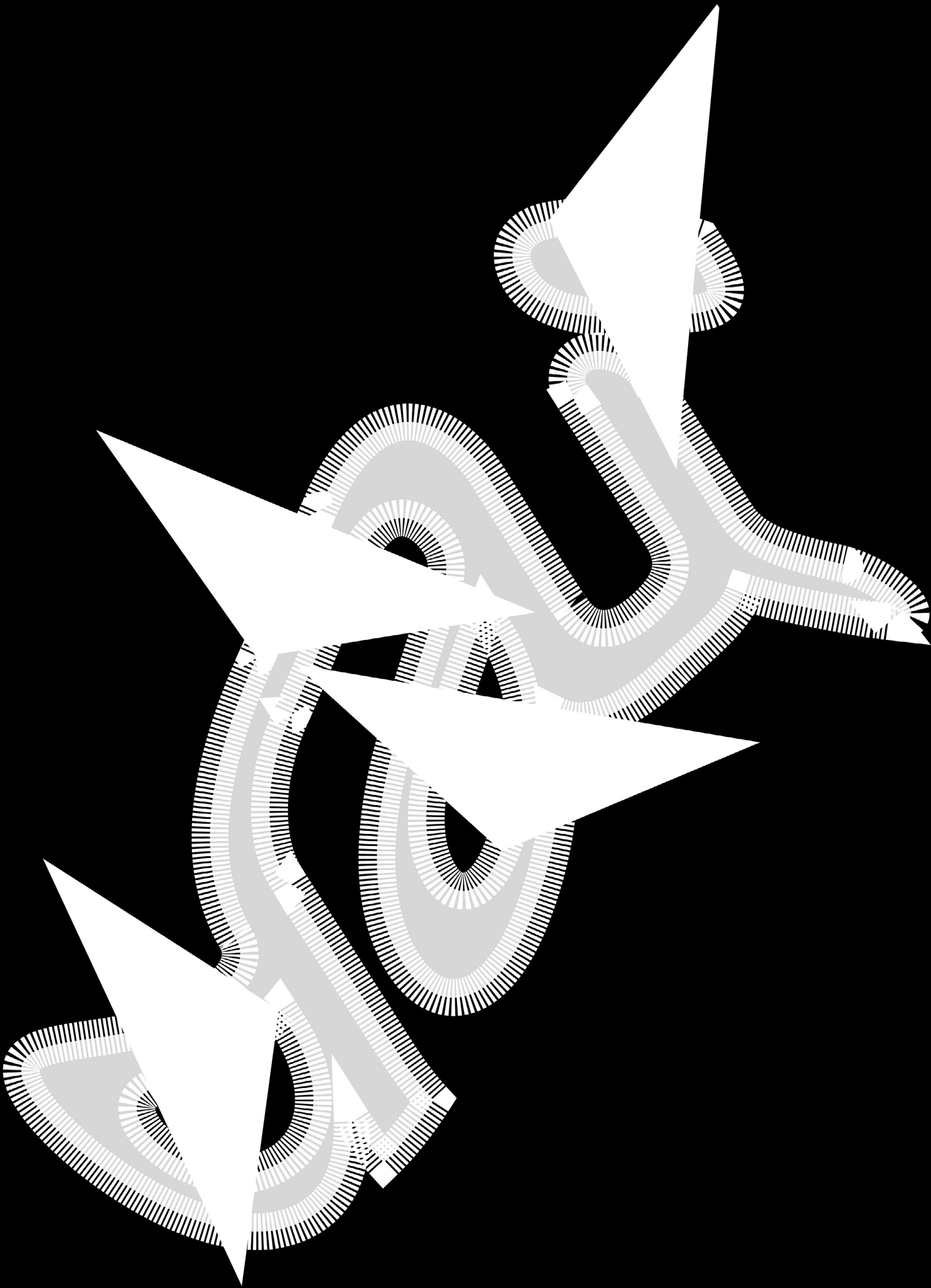
Organising committee of the dai digital
Marc Engenhart
Sebastian Loewe
Julian Pohl (assistant)

Editorial design
Engenhart ° Design Studio
engenhart.design

INITIATIVE FOR APPLIED ARTIFICIAL INTELLIGENCE

DIGITAL PRODUCT SCHOOL

Table of contents



Learning to Educate Machines

Patrick Hebron

Research

This is the companion publication to Hebron's opening keynote address for the 2020 dai digital. The keynote can be accessed at https://www.designing-artificial-intelligence.eu.

ABSTRACT

While machine learning has drawn influence from a diverse range of fields such as biology and thermodynamics, its connection to the fields of education and anthropology have been dramatically under-explored. In the human world, we understand that the intellectual capacities of the individual cannot exist in a vacuum. Rather, they are built upon the transference of distilled knowledge and our relation to other individuals who help us to notice the world and reflect upon our actions within it. In the machine world, our tools for shaping the learner's experience are still woefully crude. In this paper, we will look at teaching and curriculum design as creative acts, essential to the advancement of artificial intelligence and to the furtherance of the creative partnership between humans and machines.

1. SEEING THE OTHER

We are inclined to imagine artificial intelligence (AI) through an anthropomorphic lens. This is understandable. We have no better example of intelligence than ourselves and therefore view human intelligence as defining 'intelligence' in general. But there is much to be learned about the external world and about ourselves by understanding other beings through their own lens—by not conforming them to our image of ourselves. As Jean-Paul Sartre says, "the Other is the indispensable mediator between myself and me."[1] That is, we cannot truly understand our own intelligence without contrast.

Admittedly, there are already other greatly intelligent creatures in our midst—primates, dolphins, elephants, octopi and so forth. But their worlds and the problems they face may be too different from ours (and in other ways too similar) for us to truly make sense of ourselves through the contrastive lens of these species. The other provides a kind of shortcut for discovering alternate ways of being in and navigating the world. In Reinforcement Learning, there is talk of the Explore vs Exploit tradeoff—look for a new strategy or continue using one that has already proven successful. By observing the other, we can sometimes glean successful strategies without the labor of exploring for ourselves. The other's strategy may utilize very different premises or facts about the environment from those we have encountered in our own explorations. If our own experiences have not led us to those same facts or if the strategy's successful outcome cannot be immediately observed, then we are unlikely to discern the value of the strategy a priori.

In AlphaGo's famous "Move 37," the commentators thought the move was a mistake. The value of the move was ultimately confirmed by the machine's win and led Lee Sedol and other Go

1 Jean-Paul Sartre, *Being and Nothingness* (New York: Washington Square Press, 1992), 302.

experts to re-evaluate their understanding of the game. But if the machine happened to lose that particular game, the potential value of the strategic approach underlying this move should not necessarily be discarded. The educational philosopher, Patricia Carini, said:

> I have to trust that what I am attending to makes sense; that it is not a merely accidental or chance event. To discover the subject's coherence and how it persists in the world, I have deliberately to shift my own perspective in relation to it.[2]

Seeing value in the other's approach before it is confirmed requires open-mindedness. There is an opportunity cost to entertaining alternate strategies. Outside of virtual environments, it is often the case that the scenario can only be played-out once—making it impossible to compare divergent strategies. Why would anyone ever take a chance on a strategy that does not jive with one's own intuition and experience?

It would be reasonable to conclude that observing the other is simply another mode of exploration, which requires time commitment without guaranteed reward. This is true, but even without that confirmation, even if the other's approach turns out to be flawed, it gives new vantage points on how to be in the world. It still involves a kind of exploration, but it is an exploration of the world already mediated, already trampled through. In this sense, it gives us new premises and vocabulary that we may be unlikely to uncover ourselves. It may redirect us from the paths of exploration to which we are otherwise inclined.

In the human realm, a certain emotional availability to the prospect of learning through the observation of other humans comes from the fact that they are human. We see a "whole self" behind the work of another human and we put trust in the likeness of that self to our own—the experiences and properties they presumably share.

The AI is not a whole self, or at least not one in the sense to which we are accustomed. An anthropomorphic presentation of the AI helps us to project a sense of wholeness upon it, thereby making it a more palatable subject from which to learn.

The AI is a specialist. But in some key respects, we are specialists too. We would not be able to pilot an octopus body. Similarly, an artificial general intelligence will have a specific domain of experience, specific sensors and actuators, even if it can reason in a general way. But those experiences and capabilities will be different from ours. How can we learn to see value in them? How can we learn to use them in our own growth?

The educational theorist Howard Gardner offered the "Theory of Multiple Intelligences," which differentiates human intelligence into specific 'modalities,' rather than seeing intelligence as a single general ability. Despite the flaws and limitations of this theory, through a humanistic lens, we may see this work as trying to show value in the effort of students whose areas of specialization may not be otherwise valued by our educational system. Kofi Annan said:

2 Patricia Carini, *Starting Strong: A Different Look at Children, Schools, and Standards* (New York: Teachers College Press, 2001), 163.

> Never walk into a situation believing you know better than the natives. Keep an open mind. You have to listen and look at how they do things. Otherwise you can make some very serious mistakes. [3]

There is much to be learned from a specialist operating in a complex domain with which we have limited familiarity. Of course, the funny thing about these AI specialists, these "natives," is that they did not really land on the shores of the specialized domain before we did. We pointed them to it and determined which aspects of it would be visible to them. So, this endeavor requires something from both the human and the AI. We also determine much about the environment and experiences from which our human children learn. But it is ultimately they who must do the learning. Once we set them upon the shores, they may be able to spend more time there than we could ever hope to ourselves. It will become theirs and if we set them up to explore it fully, then we may benefit from the time they spend there.

3 Kofi Annan, "A Champion for Development, Security and Human Rights," interview by Elaina Loveland, *International Educator*, (May-June 2013): 24.

4 Geoffrey Hinton, source unknown.

2. CREATION AND INFLUENCE

As Geoff Hinton said, "If I want to change the ideas in your head, I can't just reach in and tweak what the neurons are doing. I have to change the experiences you learn from."[4] To put it another way, intelligence cannot be created, it can only be influenced.

As much as we think we teach human children to do math, we can really only set up the right circumstances for them to learn it themselves. Even if we can teach arithmetic by rote memorization, human progress by definition depends upon the next generation advancing beyond the knowledge of the previous. As such, we need to cede control. We need to stop teaching and instead let learning happen. We do this by constructing the right environment for learning to occur.

From this point of view, I believe the future of machine intelligence has more to do with design than with engineering. Engineering, or at least good engineering, means precisely controlling the behaviors of a system. Design, or at least good design, means creating a system that enables the user to go beyond the original intent of the designer—a system that can be adapted, leveraged, built upon.

For machine intelligence to be of true value to our society, it must go beyond explicitly engineered behaviors and knowledge. It must learn for itself. It must exceed what we have imbued upon it. If it is to be useful to us, then we must create an environment through which the machine's exceeding of our explicit engineering will still adhere to the values and goals of its creator and of the society in which that creator is embedded.

3. HOW WE TEACH MACHINES TODAY

Though neural architectures have advanced by leaps and bounds over the last few decades, the manners in which we teach machines and design curriculums for them remain incredibly crude. To better understand this, let's take an anthropomorphic look at three forms of machine learning:

The *supervised learning* system is much like a Spelling Bee champion. It does nothing but look at examples, millions of examples, of one specific kind of information. Its experience of the world is completely lacking in diversity. It can perform spectacularly at one task, but can do nothing else. In the human realm, it is unlikely that this kind of experience leads to a truly innovative thinker.

The *unsupervised learning* system is much like a singer performing a song in a language he or she does not speak. Like the supervised system, this learner does nothing but look at millions of examples. But here, the goal is different. The system learns the internal patterns present within a set of example inputs so that it can produce very similar patterns. Yet, this ability to parrot the examples it has learned from never really leads to an understanding of what those patterns mean or purpose they serve outside of themselves.

The *reinforcement learning* system is much like a child abandoned in the wilderness. It is left entirely to its own devices and must learn through pure trial-and-error in a world that either rewards or punishes its actions. This unaided manner of learning about the world is not very efficient, but if the individual can get enough experience without getting killed, it will develop a strong intuition for navigating the world. Through the notion of a 'wolf child', we have reason to suspect that what really makes us human is not our human brains, but rather our being raised within human society, within human families.

What does a teacher do for the student? In contrast with the pure trial-and-error approach of reinforcement learning, a teacher helps to distill relevant observations that will make the learner's process more efficient. Teachers help the student to notice the world and its properties. They help the student to notice his or her own actions within the world. For example, a teacher may point out that a bike will stay upright only if it is already in motion or that the rider's knee is turned out a bit too far.

The teacher also helps the student to plot a course through increasingly complex materials. How does a person reach the ability to read academic literature and synthesize new insights from it? You cannot just put an infant in a library and expect a medical expert to emerge! The curriculum needs to be scaled up by introducing increasingly complex concepts progressively and by pointing the student to some salient, high-level observations at each stage in complexity scaling.

4 VISUALISING EXPERIENCE

I want to return our focus to the ways in which AI can help us to explore the world more effectively and to the idea that:

> The other's strategy may utilize very different premises or facts about the environment from those we have encountered in our own explorations.

This is a daunting problem, which should not be overlooked in the design of systems that task machines with accomplishing something that is to be of value to humans. In some cases, human comprehension may not be necessary. One such example is DeepMind's system for optimizing electrical grid usage. Here, the system's utility from a human point-of-view is derived purely from its outcome. If electrical consumption is reduced without interfering with human behavior (i.e. without causing blackouts), we have no real need to understand the strategies taken by the machine nor the experiences that led the machine to those strategies.

Similarly, if an AI system were to develop a successful vaccine or treatment for Covid-19, we would rejoice in and benefit from the machine's achievement, regardless of whether we understood its strategy or methodology. Of course, it might be beneficial for future virology research to be able to incorporate the machine's ideas into academic literature. But, if this were not possible, we could instead simply set ourselves on the idea that this technical discipline could advance even in the absence of human comprehension. So long as its technical goals continued to be met, we would have no real cause to require greater human comprehension.

There are many scenarios, though, in which this sort of blind faith cannot be sufficient. If the success of the system cannot be measured in purely objective terms, then we need to design systems that will make the machine's strategy more comprehendible to us—helping us to see for ourselves the machine's experiences, the premises or facts about the environment, that led the machine to those strategies.

Watching footage of the reinforcement agents in OpenAI's "Hide and Seek" gives us something more substantive than mere personification.[5] It gives us the opportunity to see and to relate to the scenarios played-out by the learning system. It enables us to not only accept the machine's conclusion, but see firsthand how it arrived there.

There is great value in the human trainer being able to observe the dataset or the training scenarios upon which the machine was trained. The difficulty in this, of course, is the volume of experience required by the machine to reach its novel capabilities. Supervised learning datasets routinely include millions of individual artifacts. Reinforcement learning systems must often perform tasks millions of times. It is simply impossible for us to be present for all of that.

5 "Emergent Tool Use from Multi-Agent Interaction," OpenAI, September 17, 2019, https://openai.com/blog/emergent-tool-use.

6 Seymour Papert, Mindstorms: Children, *Computers and Powerful Ideas* (New York: Basic Books, 1980), 20-21.

The subject of explainability in machine learning has been an increasingly prominent one. The hope, it seems, is that the machine would be able to distill its experience into some higher level (probably symbolic) form in order to convey the logic that underlies its strategy. But, it is important to note that this is often not possible even for a human expert.

When a radiologist is able to associate a particular pattern in an image with a particular diagnosis, that association comes as something much like a sensory-motor response. Having seen many positive and negative examples before, the radiologist has a kind of impulse response. The pattern simply triggers the association in the radiologist's mind. It seems to come as a spark of intuition. There may be no better explanation than a reference to the voluminous cloud of prior experiences.

Perhaps, then, we should take the stance that in many cases, explainability in AI will simply mean the system's ability to point to particularly instructive training samples or scenarios. Through a highlight reel of the machine's experience, or through the machine's being able to interactively recall a particular experience on demand for the human inquisitor, we may find our way to a shared understanding of the machine's insights.

5. THE CONSTRUCTIVISM OF OTHER INTELLIGENCES

How does better teaching of machines come back to us? How does it extend our reach? The answer is that the roles of teacher and student must form a loop. By letting the student learn and through our observation of their manner of learning, we take on a new perspective. We have to interact with the learner's learning. This brings us to our final destination—constructivism.

The great constructivist, Seymour Papert, said in his seminal book, *Mindstorms*:

> Even the best of educational television is limited to offering quantitative improvements in the kinds of learning that existed without it. „Sesame Street" might offer better and more engaging explanations than a child can get from some parents or nursery school teachers, but the child is still in the position of listening to explanations. By contrast, when a child learns to program, the process of learning is transformed. It becomes more active and self-directed. In particular, the knowledge is acquired for a recognizable personal purpose. The child does something with it. The new knowledge is a source of power and is experienced as such from the moment it begins to form in the child's mind.[6]

To interact with something is to internalize it—to map it, to graft it onto oneself. It is in this spirit that we must encounter the experience of the artificial intelligence. To learn from its novel insights, we must bring our own experiences into proximity with those that led the machine to the insights in the first place. Our tools today for building, training and visualizing the experiences of machines are hopelessly crude. Above all else, this is a design problem. As with any design problem, start from the user—in this case, there are two.

BIBLIOGRAPHY

Annan, Kofi, "A Champion for Development, Security and Human Rights," interview by Elaina Loveland, *International Educator*, (May-June 2013): 24.

Carini, Patricia, *Starting Strong: A Different Look at Children, Schools, and Standards* (New York: Teachers College Press, 2001), 163.

"Emergent Tool Use from Multi-Agent Interaction," OpenAI, September 17, 2019, https://openai.com/blog/emergent-tool-use.

Papert, Seymour, *Mindstorms: Children, Computers and Powerful Ideas* (New York: Basic Books, 1980), 20-21.

Sartre, Jean-Paul, *Being and Nothingness* (New York: Washington Square Press, 1992), 302.

www.designing-artificial-intelligence.eu

ISBN 978-3-00-068634-4