AI for Architects

Elena Gavagnin

Trying to define what "AI" is can feel like trying to catch a slimy fish that keeps slipping when gripped, leaving you with only buzzwords in your hands. The term in itself is much older than the current hype, which started in the second decade of the twenty-first century after the "data science" hype. The quest for machines to perform "intelligent" tasks, or tasks linked to the expression of some sort of intelligence, was formally recognized and named "AI" in the mid-1950s by researchers including John McCarthy, Herbert Simon, and Arthur Samuel. However, the conceptual groundwork can be traced back to Alan Turing, and even further if we consider the foundational algorithmic principles at the basis of much modern AI.

Clearly, in order to define and explain what AI is and how it works, a natural starting point is the definition of "intelligence." This, however, seems to be the critical yet crucial part in the "sliminess" of AI. One of the best definitions available is given by the very same John McCarthy and reflects the broader context in which it was formulated: "intelligence is the computational part of the ability to achieve goals in the world."[1] Intelligence is directly linked to the ability to compute something and the concept of intention. "To achieve goals" appears to be the distinguishing prerogative of an intelligent entity, even though this is an inherently relative concept, since the notion of a goal depends on the context and, in particular, on the perspective of an observer. Richard Sutton expands on this relativity, stating that "a goal-achieving system is one that is more usefully understood in terms of outcomes than in terms of mechanisms."[2] This suggests that an observer perceives such a system primarily through its outcomes rather than through the underlying mechanisms driving it.

[1] John McCarthy, "What is artificial intelligence?" (Stanford University, 2007), 2, http://www-formal.stanford.edu/jmc/whatisai.pdf.

[2] Rich Sutton, "The definition of intelligence," *Incomplete Ideas* (blog), July 9, 2016, http://incompleteideas.net/IncIdeas/DefinitionOfIntelligence.html.

The quest for intelligent systems is often, in a first approximation, reduced to the problem of "how to construct computer programs that automatically improve with experience."[3] This reflects the emergence of machine learning as a key branch of AI—one that approaches intelligence by enabling machines to learn from data and refine their performance over time, rather than relying solely on manually programmed rules. A formal and well-posed definition of learning is given by Tom Mitchell: "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."[4]

This definition broadly defines learning as any automatic improvement connected with experience. Historically, however, three distinct paradigms of learning have emerged: supervised, unsupervised, and reinforcement learning. Apart from involving different tasks and performance metrics, their main underlying distinction lies in the amount of human supervision required, which implicitly affects how and what kind of experience the system acquires.

Supervised learning assumes that a model's predictions can be verified against a human-defined ground truth, similar to checking answers at the back of a math textbook after solving a problem. The system compares its predictions with the correct answers, which must be explicitly provided.[5] In this case, experience consists of a collection of labelled examples—that is, data accompanied by their correct outputs. In contrast, unsupervised learning involves discovering patterns and structures in data without labels by analyzing distributions and relationships within the dataset.[6] Here, experience is represented by the amount of unlabeled data, which enables the detection of hidden structures and meaningful groupings. The third classic learning paradigm is reinforcement learning, in which an agent learns through trial and error, updating its behavior based on feedback from the environment.[7] In this case, experience is represented by the agent's interactions with the environment, where it takes actions, observes new states, and

[3] Tom Mitchell, *Machine Learning* (McGraw-Hill, 1997), xv.

[4] Mitchell, *Machine Learning*, 2.

[5] Mitchell, *Machine Learning*, 2; Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, 2009).

[6] Christopher Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).

[7] Richard S. Sutton and Andrew Barto, *Reinforcement Learning: An Introduction* (MIT Press, 1998).

receives rewards or penalties. The more it interacts, the more it learns and improves. These three classic learning modalities closely resemble how humans learn. We can intuitively recognize parallels with instruction-based learning,[8] discovery/statistical learning,[9] and operant conditioning.[10] However, in humans, these learning modes are not strictly separate—they often blend, overlap, and influence each other in complex ways.

In all cases, we have seen that the experience is represented by the data available, whether in the form of examples or interactions with an environment. The need for data to enable learning naturally leads to the necessity of collecting and representing these data. Recording as much data as possible becomes central for AI, hence the well-known hype around Big Data & Co. The fact that learning requires data is not surprising—after all, the same applies to humans—but AI models require a strictly numerical representation of information to perform computations. By having to define a way to "encode"—to represent—information numerically, it becomes clear that the choice of representation also influences the effectiveness of learning.

One major shift in AI research has been the recognition that the way data are represented internally by a system significantly affects its ability to generalize and transfer knowledge. Representational learning then emerged as a paradigm for automatically deriving abstract features.[11] Rather than manually engineering which edges, shapes, or keywords are important, the system learns a latent space that captures the intrinsic salient patterns. A latent space is a compressed, abstract representation of data in a lower-dimensional space. One form of representational learning is self-supervision, in which models learn general features from raw data without human-provided labels. This is made possible by designing so-called pretext tasks, where the "labels"

[8] Robert M. Gagné, *The Conditions of Learning and Theory of Instruction* (Holt, Rinehart & Winston, 1965).

[9] Jerome S. Bruner, "The act of discovery," *Harvard Educational Review* 31, no. 1 (1961): 21–32; Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport, "Statistical learning by 8-month-old infants," *Science* 274 (1996): 1926–28.

[10] B. F. Skinner, *The Behavior of Organisms: An Experimental Analysis* (Appleton-Century-Crofts, 1938).

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature* 521 (2015): 436–44, https://doi.org/10.1038/nature14539.

come automatically from the structure of the data itself. Autoencoders[12] can be considered an early approach in this direction: A network tries to reconstruct its own input, effectively creating “labels” from the input itself. In modern deep learning, language modeling (predicting the next word) also serves as a self-supervised objective, since the data themselves provide the target (the “next word”). Other examples of pretext tasks include predicting the context of an image patch,[13] colorizing a grayscale image,[14] or reassembling jigsaw puzzles.[15] Representational learning represents a further step in that not only are the model’s parameters learned, but also the input features themselves.

Interestingly, the concept of representational learning finds an evocative, if more artistic, parallel in the Dear Data project.[16]

In this collaborative experiment, two designers collected personal data on aspects of their daily lives—ranging from coffee consumption to emotional states—and transformed those raw measurements into hand-drawn visualizations. Although not algorithmic, Dear Data demonstrates how the process of deciding what to collect and how to depict it can provide emergent insights. In AI-based representational learning, a similar but automated process unfolds at scale: data of various forms (images, text, sensor readings) are mapped onto multi-dimensional latent spaces that encode higher-level semantics within the data.

One of the advantages of having an abstract, multi-dimensional latent representation of a data point (e.g., an image) is that it transforms the original form (an array of pixels) into a more general encoding that captures distinguishing features and contextual meaning. Such a general representation be-

[12] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, “Learning representations by back-propagating errors,” *Nature* 323 (1986): 533–36, https://doi.org/10.1038/323533a0.

[13] Carl Doersch, Abhinav Gupta, and Alexei A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015), 1422–30, https://link.springer.com/chapter/10.1007/978-3-319-46466-4_5.

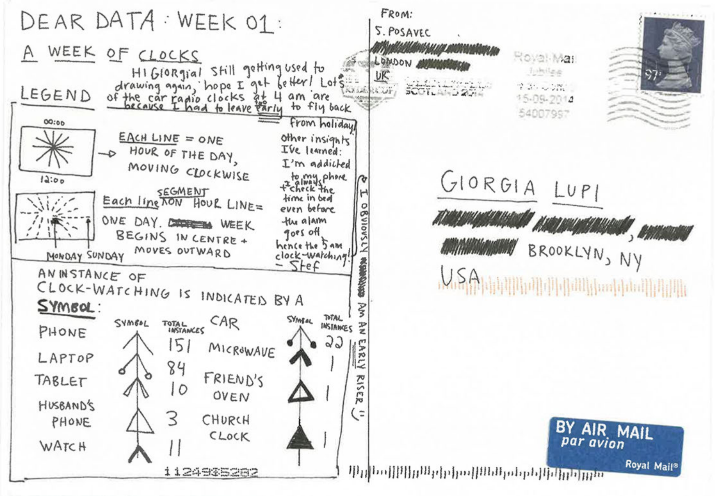
[14] Richard Zhang, Phillip Isola, and Alexei A. Efros, “Colorful image colorization,” in *European Conference on Computer Vision (ECCV)* (2016), 649–66, https://link.springer.com/chapter/10.1007/978-3-319-46487-9_40.

[15] Mehdi Noroozi and Paolo Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *European Conference on Computer Vision (ECCV)* (2016), 69–84, https://link.springer.com/chapter/10.1007/978-3-319-46466-4_5.

[16] Giorgia Lupi and Stefanie Posavec, *Dear Data* (Princeton Architectural Press, 2016).

comes universal and transcends data-domain boundaries—such as text versus visual data—since these representations can be put in mutual relation and jointly trained. Contrastive multimodal learning models (e.g., CLIP[17]) operate on multiple data modalities within this shared latent space, enabling them to learn representations that capture semantic parallels across different formats. As a result, they can generate captions for images or produce images from text. Another peculiar aspect is that in latent spaces learned by neural networks (especially in large-scale models), semantically-related items tend to be close together. Sampling near a known point produces outputs that share semantic meaning or structural appearance qualities, which is the idea at the basis of generative AI.

Fig. 4: Giorgia Lupi & Stefanie Posavec, Dear Data, 2016



DEAR DATA : WEEK 01 :

A WEEK OF CLOCKS

HI GIORGIA! Still getting used to drawing again, hope I get better! Lots of the car radio clocks at 4 am are because I had to leave too early to fly back from holiday!

LEGEND

00:00

12:00

EACH LINE = ONE HOUR OF THE DAY, MOVING CLOCKWISE

Each line SEGMENT ON HOUR LINE = ONE DAY. WEEK BEGINS IN CENTRE + MOVES OUTWARD

MONDAY SUNDAY

Other insights I've learned: I'm addicted to my phone + I always check the time in bed even before the alarm goes off, hence the 5 am clock-watching! – Stef

I OBVIOUSLY AM AN EARLY RISER :-)

AN INSTANCE OF CLOCK-WATCHING IS INDICATED BY A SYMBOL:

	SYMBOL	TOTAL INSTANCES		SYMBOL	TOTAL INSTANCES
PHONE		151	CAR		22
LAPTOP		84	MICROWAVE		1
TABLET		10	FRIEND'S OVEN		1
HUSBAND'S PHONE		3	CHURCH CLOCK		1
WATCH		11			

i1249352B2

FROM:
S. POSAVEC
[redacted]
LONDON [redacted]
UK

SCOTLAND 2014

Royal Mail
15-09-2014
54007997

97p

GIORGIA LUPI
[redacted],
[redacted] BROOKLYN, NY
USA

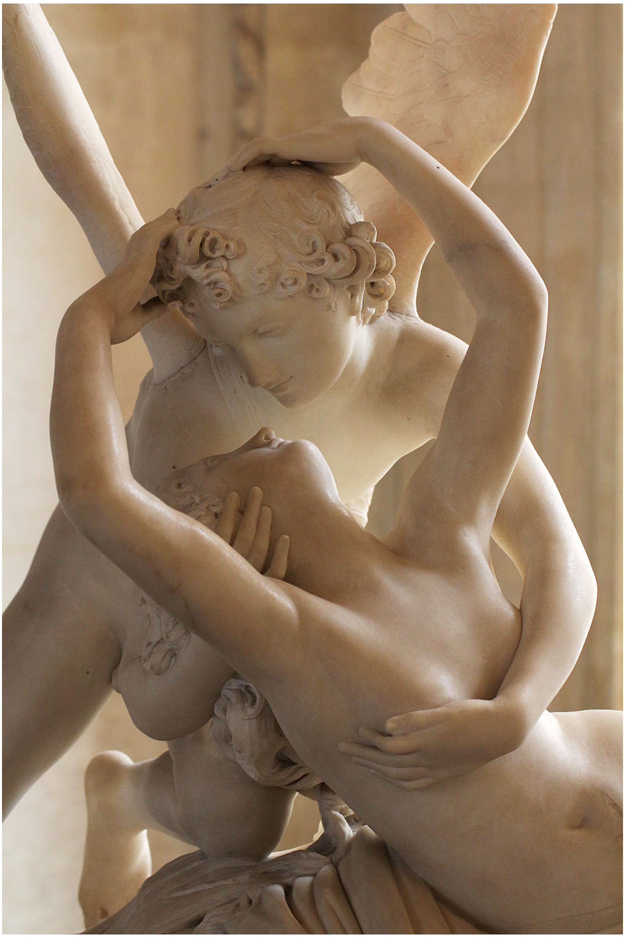
BY AIR MAIL
par avion
Royal Mail®

The euphoric advancements and remarkable successes of the described approaches—originally driven by the quest for intelligence and the ambition to build learning machines—can give the impression that modern AI is capable of almost anything. However, this is not (yet) the case. Asked, for example, to

[17] Alec Radford et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)* (2021), https://arxiv.org/abs/2103.00020.

distinguish which hand rests on Psyche's head in Canova's *Psyche Revived by Cupid's Kiss*, most visual-language models fail to provide the correct answer—a situation that, until recently, also applied to Leonardo's *Mona Lisa*.

Fig. 5: Antonio Canova, Psyche Revived by Cupid's Kiss, 1787



When faced with the same question, humans naturally answer from the perspective of the person whose hand it is, not from that of the observer. In other words, the right hand of a person is never referred to as the left one just because it appears on the left side from an observer's point of view. This, however, is not always the case for AI models. The surprising shortcomings in tasks that come naturally to humans—such as social and spatial cognition, especially perspective-taking—highlight an important gap. This ability, deeply connected to our innate sense of physical presence, our embodied experience of space, and our awareness that others have bodies both different from and similar to our own, invites deeper reflection on AI within an architectural context.

What does it truly mean to feel the space and immerse ourselves in it as well as in the people around us? And is this a prerequisite to achieve truly intelligent machines?