An Unconventional Look at AI: Why Today's Machine Learning Systems are not Intelligent

Nancy Salay[1]

Machine learning systems (MLS) that model low-level processes are the cornerstones of current AI systems. These 'indirect' learners are good at classifying kinds that are distinguished solely by their manifest physical properties. But the more a kind is a function of spatio-temporally extended properties — words, situation-types, social norms — the less likely an MLS will be able to track it. Systems that can interact with objects at the individual level, on the other hand, and that can sustain this interaction, can learn responses to increasingly abstract properties, including representational ones. This representational capacity, arguably the mark of intelligence, then, is not available to current MLS's.

Introduction

The current rhetoric has it that AI is here, or, at least, around the corner. But if this claim is false, then treating systems as such, that is, relying on them to make judgements, could have grave consequences. Autonomous vehicles are just one example of the many new AI technologies poised to enter the public space. Since stemming this technological tide seems futile, the academy has a responsibility to raise the public's understanding around what constitutes intelligence. Here I begin this effort by arguing that we should stop thinking about current machine learning systems (MLS's) as 'intelligences' since they do not model the learning necessary for intelligent behaviour.

What is Intelligent Behaviour?

The very first question we might ask, then, is "What is it to act intelligently?" Examples help orient our thinking so let us begin with a few here. A robot that has been programmed to perform basic household chores, but which freezes when confronted with an unfamiliar task or situation, is not behaving intelligently: it can do only what it has been programmed to do. Similarly, any machine that does only what the limits of its design permit and no more — *calculators*, *toasters*, *cranes* — does so without what we would call intelligence. In contrast, beings who make novel responses to novel situations, who 'figure out' what to do next, provide us with paradigmatic intelligent behaviour. Think here of the battery of researchers currently working on the problem of developing a Covid-19 vaccine. The task draws on much background knowledge, to be sure, but it also requires a capacity for solving new kinds of problems. If we were not confident that these scientists were capable of such 'out-of-the-box' thinking, we'd have no reason for hope.

But we need to be careful — a novel response is not always an indicator of an intelligent response. Many animals are capable of adapting to changing situations in what seem like 'pre-programmed' ways: viruses mutate when hosts become resistant; cockroaches change food sources during resource scarcity; and, primates shift their group dynamics when territories diminish. We call the capacity for this sort of behavioural change 'adaptability.' While there is a relationship between adaptability and intelligence — they are both capacities for learning new responses to new situations — the former is on an evolutionary time-scale, one that extends across individuals to species, and the latter is on a developmental time-scale, one that extends across an individual over its own lifetime. From the perspective of the individual, adaptive responses are canned responses, backup strategies that kick in when first pass responses fail; only from the perspective of the species are such responses

Fig. 1: Corona Virus Researchers
(Source: https://www.wired.com/story/coronavirus-research-preprint-servers/)

novel. Intelligent responses, in contrast, are local strategies that individuals develop in response to new, locally experienced, situations.

Intelligent Action is not Continual

Notice that having a capacity for intelligent action does not entail that one uses this capacity all the time. That it does was the working assumption in the early days of AI — 'Good-Old-Fashioned-AI' (GOFAI). On those completely top-down models, every action (limited to simple screen outputs in most cases since GOFAI systems were not equipped with bodies) was the product of a line of 'reasoning,' a decision. Today, thanks to the insights of Embodied Cognitive Science and, more broadly, Phenomenology, we understand that even intelligent individuals mostly act in automatic or unconsciously directed ways. Such actions might be skillful, as when someone plays an instrument or adeptly traverses a narrow cliff-side path, but such behaviour unfolds according to learned responses to the occurrent features of an ongoing situation. When an obstacle looms, the body swivels to avoid it; when a note is played, the next note in the learned sequence is anticipated. In intelligent behaviour, in contrast, factors beyond the occurrent features of the ongoing situation influence our behaviour: a note is played, but now the anticipated next note is not played. My musical execution of Beethoven's Emperor Concerto might be skillful, but when I intersperse it with the melodic line for Happy Birthday, because I know that today is yours, I have acted intelligently as well. That it is your birthday is not a physical fact in the occurrent situation in which we are participating and yet, as intelligent beings, it is something to which we are both capable of responding.

Representation makes Intelligent Behaviour Possible

How do we become responsive to such 'offline' factors? We represent them to ourselves and one another. A classic demonstration of the dramatic behavioural effect this capacity for representation can have is the infamous marshmallow test.[2] A subject, usually a child, is presented with a single marshmallow. He is

Fig. 2: Subject in Marshmallow Test (Source: https://www.cbc.ca/radio/thecurrent/the-current-for-july-8-2015-1.3142634/marshmallow-test-proves-self-control-can-be-learned-1.3142668)

told that he is welcome to eat it but, if he can sit patiently for X minutes without tasting the marshmallow, he will receive another marshmallow in addition to the first. He will receive one marshmallow if it is eaten right away and two marshmallows if he waits until the experimenter returns. Young children find it very difficult to resist the sensory draw of the marshmallow and generally give in and eat it before the experimenter returns. As children mature, however, they are increasingly able to wait for the arrival of the second marshmallow. They are capable of suffering through short term deprivation — not tasting the marshmallow that is present — for the sake of increased future gain — two marshmallows. To repeatedly not succumb to the marshmallow temptation, to what is sensorily present, children must behave now in a way that takes into account factors that are not spatio-temporally present, namely that there is a potential for future gain. Younger children, perhaps because they have not developed the relevant linguistic representational skills, are capable only of responding to the occurrent, sensory factors of the situation. Being thus completely in the moment, they gobble up the sweet-smelling treat. Another version of the marshmallow experiment,[3] this time a reverse contingency test with chimpanzee subjects, demonstrates the critical role of representations in intelligent behaviour even more clearly. A chimpanzee is presented with two plates of treats, one having visibly more than the other, and is invited to choose one of them. A second chimp sits by, observing. The plate chosen is given to the second chimp and the chooser is awarded the one not selected. Similarly to young children, though they understand the terms of the offer perfectly well, chimpanzees are incapable of resisting the overwhelming sensory draw of the treats — *their smell*, *touch*, *appearance* — and they invariably choose the plate with the greatest amount. Repeatedly they watch, furiously, as the larger pile of candies goes to the lucky bystander. But when a slightly altered version of the experiment was run with a chimpanzee — *Sheba* — who already knew some rudimentary mathematical symbols, the results were very different. Instead of being asked to choose between two plates with visible piles of candies, Sheba was offered two plates with lids labelled with the number of treats underneath. Since numbers have no sensory draw whatever, Sheba was now able to consistently make the self-maximising choice of the plate with fewer candies. By using the numbers, representing the possibility of future treats, Sheba was able to take into consideration factors that were not spatio-temporally present and thereby make the intelligent choice.

Representations have Unusual Properties

A word needs to be said here about representations themselves since they are peculiar things. A representation — for example, a sentence utterance, a physical token such as a game playing piece, or an occurrent thought — has, in addition to the usual physical properties that

Fig. 3: School Crossing Guard (Source: https://www.mlive.com/news/2012/01/traffic_talk_how_much_authorit.html)

all physical things have, representational properties that extend beyond these. My arm is a configuration of cells and impulses, and this is all that it is, nothing more than these physical properties. But the sentence token, "My arm is a configuration of cells and impulses," has the physical properties constituted by the ink on this page, the molecules in the paper, and so on, as well as the property of being about my arm. It represents the world, with respect to my arm at least, as being a certain way. The technical term for this property of representation is 'intentionality:' anything that is about/represents something beyond itself in this way is an intentional thing. All signs,[4] then, are intentional objects: words and sentences; numbers; icons. And if we think that thoughts are internal representations of the way the world is or could be, then they are intentional objects as well.

Unlike physical properties, however, representational/intentional properties are not intrinsic to things. A stop sign is a symbol in the context of road traffic systems, but outside of these contexts, to a squirrel for example, the stop sign is merely a physical thing in the world, representing nothing. Likewise, the sentence "My arm is a configuration of cells and impulses," to someone who does not speak English or to something not capable of speaking at all is just the physical thing that it is, the letter-shaped ink patterns on the page, representing nothing. Physical objects become symbols only in the context of a larger system within which individuals respond to them in ways that point beyond themselves. Such individuals are themselves intentional beings since they can respond to a symbol's representational properties as well as its physical ones. Intentionality, then, is a key aspect of intelligence.

To build an intentional machine, namely one that can use representations to guide its behaviour, is one of the central tasks of AI. The foundational assumption that continues to drive research in the field, and which has spawned the current zeitgeist of mainstream cognitive science more generally, is this: intentionality is (exhaustively) reducible to low-level processes. If we want to understand the intentional capacity of human beings, we need to look inside human beings, at their neural circuitry mostly, to see which aspects of it are responsible for intentional behaviour. In other words, the capacity to respond to a symbol's representational properties (and not just its physical ones) is nothing more than a (possibly very complex) combination of the low-level processes that constitute the response behaviour.

The problem with this idea, however, is that it is wrong: low-level processes and personal-level intentional behaviour are not comparable, and hence not identifiable, activities. Yes, of course, the sub-personal processes of Fred taken together are necessary for Fred to learn a new language, but, they are not sufficient: factors that are external to Fred, *e.g.* the way that symbols are used in his linguistic community, determine whether and how Fred's learned responses are about anything at all, which is to say, whether they are intentional at all. Furthermore, a model of just the sub-personal learning part of Fred's linguistic skill, placed in the appropriate environment, will not yield the appropriate intentional behaviour. Without coordinated, personal-level activity, Fred's neural activity is just a series of low-level processes, not linguistic at all. In other words, intentional behaviour is a personal-level activity, not a low-level one. This is not a distinction that is often made in cognitive science, but it is a critical one in the context of artificial intelligence research. If intentionality is not exhaustively accounted for by sub-personal activity, then AI systems that only model sub-personal activity, which is precisely what current MLS's do, will not be intelligent. Here I will try, in broad brush strokes, to explain how and why low-level processes cannot constitute intentional behaviour.

Low-Level Processing is not Intentional

MLS's model, albeit simplistically, the low-level neural processes that underwrite organism learning. Today's MLS's have achieved much success in classification, in 'learning' to distinguish between different kinds of objects, pictures of cats and dogs, for example. I call this 'learning' in scare quotes because the relevant classifying behaviour is achieved only indirectly, by way of the pixel-level features that con-

sistently corelate with object-level kinds such as dogs and cats. That there really are natural kinds in the world — individuals that share a statistically significant sub-set of features with other individuals — is what makes this kind of indirect learning useful. If the world were not regular in this way, if there were no natural kinds, such an approach would be useless. Imagine a world in which low-level patterns did not correspond to anything useful at the personal level, where a low-level sequence meant CAT one day and then DOG the next! This is an important factor to keep in mind: the learning success of indirect classifiers is partly determined by the high-level homogeneity of environmental conditions.

But given that our world *is* populated by dependable, statistical regularities, isn't this indirect learning good enough? For simple applications, in the context of web searches for example, it might be. As a model of personal-level behaviour, however, it falls short: there is a critical granularity gap between low-level processing and object-level action. At the low-level, interaction is with sensory bits — in the case of our example MLS these are pixels — not with medium-sized objects such as cats and dogs. One low-level processing time-step does not register at the object-level of granularity at all and, conversely, a single action at the object-level corresponds to thousands of low-level processing steps. Relative to the object-level, processing occurs at speeds that do not register as activity at all; while, relative to the processing level, object-level change occurs so slowly that, again, it does not register as change at all during a single, low-level time-step. The two levels of granularity are thus spatio-temporally distinct. Since MLS's cannot interact directly with medium-sized objects, but only indirectly over many low-level time-steps and by way of regular, low-level features, the object 'learning' they achieve will always be brittle. Change one lower-level feature of an object that the MLS has not had training on, and it breaks. For this reason, no matter how robust the low-level training is, MLS's will always be subject to adversarial attack. But, more saliently with respect to the question of intelligence, such systems could never learn to respond to an object's representational properties. More on this momentarily. The only

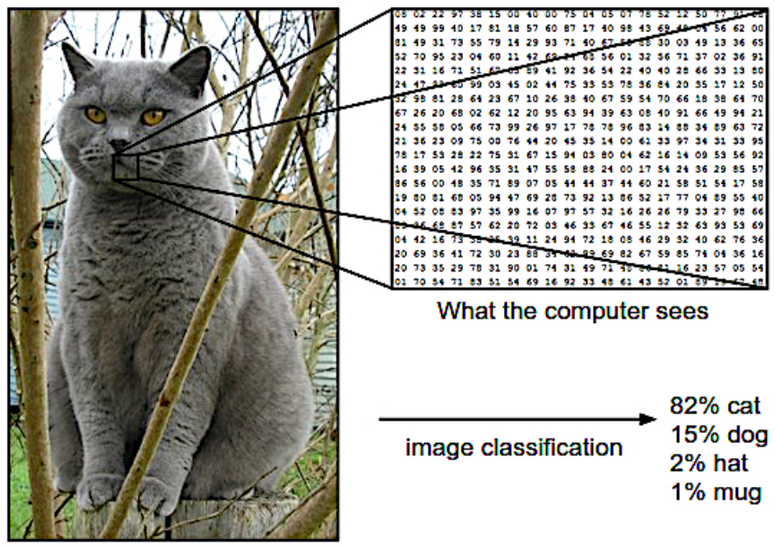

Fig. 4: Cat Classifier
(Source: https://www.kdnuggets.com/2019/11/deep-learning-image-classification-less-data.html)

sense in which a cat/dog classifier represents a distinction between the concepts CAT and DOG is from our spatio-temporally extended vantage point relative to the network, namely the vantage point of the personal-level, from which we can see that the ongoing activity of the network corresponds to a distinction between cats and dogs. From the classifier's vantage point, however, there are only ever pixels and responses to them; there are no cats and dogs at all.

To help clarify what it is for an individual to interact and learn about an object directly, let us first consider another personal-level activity: locomotion. A cat jumps from the ground to a branch by virtue of a multitude of ongoing, low-level processes that constitute its personal-level activity, but none of the low-level processes are themselves locomotions. Locomotion is something that organism wholes do. A single locomotive step, so to speak, spans a multitude of low-level processes and a single, low-level process does not map to anything at all that would count as a step at the personal-level. Some machines locomote as well. When a car moves down the road, it is the car whole that is moving, not its parts. The parts, working together, make the car locomotion possible, but there is no part or set of sub-parts that is doing the locomoting.

Perception, which is the method by which organisms interact with their environment, is likewise a personal-level activity. As with locomotion, myriad low-level (sensory) processes underwrite perceptual activity, but perception itself is a personal-level interaction with the objects in an individual's environment. For humans these are mostly medium-sized objects such as cats, dogs, tables, computers, and the like. Of course, organisms that are capable of perception do not always and only navigate their environment with perception; rather, there is an ongoing, dynamic shifting between unconscious/indirect sensory tracking and conscious/direct perception, so rapid that most perceiving individuals are themselves completely unaware of the oscillation between modes of interaction. The frequency of this oscillation, however, is a critical factor in an individual's capacity to learn to respond to the representational properties of objects since the degree to which an individual is capable of sustaining perception — of maintaining ongoing direct interaction with an object — will determine the degree to which that object can become a symbol for an individual. Before we can see how and why this is the case, a few words need to be said about symbols themselves.

Symbol Abstractness is an Indicator of Intentionality

As we have seen, a sign is anything that represents something to an agent. The more removed a sign's representational properties are from its physical ones, the more symbolic it is, while the less distinct a sign's representational properties are from its physical ones, the more natural a sign is. Natural signs and human language lie at opposite ends of this abstractness continuum. Smoke, for example, is a physical by-product of fire, serving as a natural sign of fire to any individual capable of tracking it. Many animals that already have an avoidance response to fire learn, through

Fig. 5: Forest Fire in Oregon, August 2020
(Source: https://www.oregonlive.com/pacific-northwest-news/2020/08/white-river-fire-near-mount-hood-covers-14k-acres-new-blaze-prompts-evacuations-west-of-eugene-wildfire-roundup.html)

association, to exhibit the same avoidance response to smoke. Because smoke is sensorily more dispersed than fire, and so can be perceived more quickly, it is a useful natural sign. As we move along the sign abstraction continuum, however, representational and physical properties increasingly diverge. The physical properties of the *utterance* 'Fire!' for example, are entirely distinct from its representational properties. Learning the latter requires not only a series of learning situation experiences in which different utterances are associated with different fire situations, but it also requires an entire system of symbol use within which utterances evoke appropriate fire responses. It requires a community wide practice of linguistic use. Because these representational properties extend to the features of the situations within which sign vehicles and objects occur — a child learns the word 'ball' as a result of myriad and various BALL situations — learning a response to them requires sustained interaction with the relevant sign objects, long enough for the situation to unfold. And there's the rub. MLS's track objects only indirectly, they never interact with them at all; consequently, they will not be able to sustain interaction with objects, at least not across the sorts of complex, varied situations in which humans learn. Remember that from the vantage point of the cat/dog classifier, there are no cats and dogs, only pixels. And the more abstract a sign is, the more its representational properties will trade on spatio-temporally diverse situational features rather than on a sign vehicle's physical ones.

In theory, of course, MLS's could be trained to indirectly track situations, just as they do objects. Newer deep learning models attempt to just that.[5] But success will be elusive: the properties of situations exponentially out-number those of objects, even supposing that there are statistically reliable patterns that underwrite them. In most cases there are not. As Ludwig Wittgenstein famously observed,[6] even seemingly regular situation-types such as games are impossible to pin down: some games such as baseball and soccer involve many players, others such as chess and tennis just two, and some such as solitaire only one; some games have win/lose conditions, while others — many new board games for example — are cooperative; some have fixed rules; others — *e.g.* 'playing house' — do not, and so on. Thus, the possibility for tracking even a simple type of situation such as game playing by tracking the low-level properties of the objects that might figure in them seems very small. What chance, then, is there for indirectly tracking complex, language-learning situations?

Systems such as MLS's, then, that are capable only of indirect object tracking by way of low-level properties could learn responses to natural signs but will not be capable of more abstract symbolic responses. And it is arguable that since the symbolic properties of such natural signs trade completely on their physical ones, individuals that are capable only of learning to respond to them, and not to more abstract symbols, are not intentional agents. Most new cars today are outfitted with an array of sensors that track the objects in a car's immediate surround. When my car is in reverse, for example, it will emit different sequences of 'beeps' according to how close a potential obstacle is to the rear fender. In one sense, this seems like an intentional action since it looks like the car is responding to what things in the environment signify — potential obstacle — rather than to them directly as physical objects. But, of course, the car is not beeping because it perceives a potential obstacle; rather, the triggering of one set of sensors triggers another set that in turn triggers the beep mechanism. The car has been designed so that its parts work in concert with one another in such a way that the car signals precisely when a potential obstacle is present. This is clever design to be sure, but the car itself is behaving in the sort of unintelligent, pre-programmed way we began this discussion with. At best we can say that it has been designed to respond to a natural sign.

Conclusion

Because they cannot interact directly at the object-level, current MLS-based systems are not capable of the sustained interaction with objects required for developing responses to abstract symbols. As they are fundamentally non-intentional entities, then, we should not treat them as intelligent systems, no matter how clever their design. Does this mean that there is no possibility of creating an AI system, perhaps out of a complex configuration of these networks? No. But a potential AI will need a capacity analogous to basic perception, namely, a capacity for direct interaction with the objects in its environment. In the case of humans this is sentience, sometimes called 'pre-reflective consciousness:' our basic capacity to experience our world, not simply infer it. But we still need a better understanding of what this is before we can start building systems that exhibit it.

[1] Associate Professor at the Department of Philosophy, Queen's University, Canada.

[2] W. Mischel & E. B. Ebbesen, "Attention in delay of gratification," *Journal of Personality and Social Psychology 16 (2)*, 1970, 329-337.

[3] S. T. Boysen, G. Bernston, M. Hannan, and J. Cacioppo, "Quantity-based inference and symbolic representation in chimpanzees (Pan troglodytes)," *Journal of Experimental Psychology: Animal Behavior Processes 22*, 1996, 76-86.

[4] 'Sign' is a technical term for a unit of meaning.

[5] J. Chen, K. Li, Q. Deng, K. Li, S. Y. Philip, "Distributed deep learning model for intelligent video surveillance systems with edge computing," *IEEE Transactions on Industrial Informatics*, 2019, 1-8.

[6] L. Wittgenstein, *Philosophical Investigations*., trans. G. E. Anscombe, Oxford, Blackwell, 1967, §83.