Resolute Resolutions (a blog), volume 2:
Part 1 of N: An Informal [in]Formality: AI?

January 2021

Allen Schmaltz

Abstract

We briefly consider one possible informal definition of the goal of AI, highlighting that this notion of AI has no direct connection to human intelligence.

At a coarse resolution, there are many conflicting notions of what "AI research" is, does, or seeks to achieve. On the one hand, it may very well be more productive to forgo defining a single, abstract umbrella under which all AI research falls, and simply focus on concrete engineering problems. On the other hand, abstractions may inform more philosophical considerations that we encounter in the former. In any case, here, we briefly consider one informal, abstract notion of AI. As with other abstractions, it may, or may not, turn out to be informative for downstream concrete empirical research. *We introduce it as a thought experiment, as it is a notion that involves no equivalence, nor direct relevance, to human intelligence.* We view this as a possible contrast; it is not necessarily more useful than alternatives. In future posts, we may examine alternatives, as we further consider these issues in the context of regulating and assessing real-world models.

One conceivable goal (or equivalently, "problem") of AI research is to discover the minimal program that can uncover new knowledge across environments, consistent with the observed knowledge, as understood in the following terms.[1]

In this view, an environment, E, consists of three structures, two functions, and one distribution: $< \mathbb{X}, \mathbb{Y}, \mathbb{A}, w, \mathcal{D}, w' >$.

1. A set of QUERY sequences[2], $\mathbb{X}$;
2. A set of SUPPORT sequences[3], $\mathbb{Y}$;
3. A set, $\mathbb{A}$, consisting of three atomic symbols, **U**, **T**, **F**;
4. A deterministic[4] function, $w : \mathbb{X} \times \mathbb{Y} \to \mathbb{A}$, that takes a QUERY sequence, $x \in \mathbb{X}$, and a SUPPORT sequence, $y \in \mathbb{Y}$, and indicates whether the relationship between x and y is unknown, **U**, valid, **T**, or invalid, **F**;
5. A function, $w' : \mathbb{X} \times \mathbb{Y} \to \mathbb{A}$, which emulates w but exchanges the output, **U**, **T**, **F**, according to some distribution $\mathcal{D}$. As a result, some **U** indicated by w' are underlying unknown relationships in the environment, and some are the result of artificial elision, which must be recovered via imputation. The remaining exchanges amount to adding noise to the environment.

The purpose of the artificial intelligence agent, hereafter, "program", P, which has access to $< \mathbb{X}, \mathbb{Y}, \mathbb{A}, w' >$ from E, is to serve as a replacement, possibly non-deterministic, of the underlying w of the environment, with the goal of reducing the number of relationships assigned an unknown label, while maximizing consistency of the known relationships, under the constraints across all environments, as described below. In this convention, we seek not so much to carry out particular human tasks in a particular environment, the

[1] Note that this is a blog post/thought/opinion piece.

[2] A sequence could be text, a single object, a video, a series of actions by a car, etc.

[3] A SUPPORT sequence may also include what we typically consider labels in classification settings, and in that case, would also encode the particular task itself, and/or any additional sequences to express the relationship with the QUERY. In this way, the QUERY and SUPPORT sequences are used to encode the knowledge contained in the environment.

[4] We are a bit fast and loose on this point, and it is perhaps too much specificity at this level of abstraction. For the purposes here, assume that the environment will terminate w and assume it outputs **U** if w does not terminate within a given number of steps that it specifies.

traditional domain of machine learning, although we might well get that as a side-effect of the endeavor[5], but rather we seek the minimal program that can uncover new knowledge across environments, where uncovering new knowledge is equivalent to reducing the number of unknown relationships, provided they are consistent with the known relationships across environments, as noted below.[6]

More specifically, we can describe and evaluate a program[7], P, across the universe of N environments, $\mathcal{E}$, as follows. For each environment E^i in $\mathcal{E}$, we create 2 new sequences from each $y \in \mathbb{Y}$ by concatenating the $\textbf{\textsf{T}}$ and $\textbf{\textsf{F}}$ symbols to the SUPPORT sequence: $y_t = (y, \textbf{\textsf{T}})$ and $y_f = (y, \textbf{\textsf{F}})$. Then we run P to create representations that summarize each environment:

1. For each (x, y) pair corresponding to $w'(x, y) = \textbf{\textsf{T}}$ in the environment, we run P over x to create a representation, $\boldsymbol{r}_x$, and over each of the corresponding y_t and y_f to create $\boldsymbol{r}_{y_t}$ and $\boldsymbol{r}_{y_f}$, respectively. In addition to creating the representations, P aims to minimize the distance, under some notion of distance, between $\boldsymbol{r}_x$ and $\boldsymbol{r}_{y_t}$ and maximize the distance between $\boldsymbol{r}_x$ and $\boldsymbol{r}_{y_f}$. In other words, we seek for P to produce faithful representations of the known relationships in the environment, as given to it under w'.

2. Analogously, we run P over each (x, y) pair corresponding to $w'(x, y) = \textbf{\textsf{F}}$ in the environment, where in this case, P aims to minimize the distance between $\boldsymbol{r}_x$ and $\boldsymbol{r}_{y_f}$ and maximize the distance between $\boldsymbol{r}_x$ and $\boldsymbol{r}_{y_t}$. This and point (1) correspond to *memorization* of the known relationships, as given by w'.

3. For each (x, y) pair corresponding to $w'(x, y) = \textbf{\textsf{U}}$ in the environment, we run P to produce $\boldsymbol{r}_x$, $\boldsymbol{r}_{y_t}$, and $\boldsymbol{r}_{y_f}$. P aims to minimize the distance from $\boldsymbol{r}_x$ to form the correct relationship and maximize the distance to the incorrect relationship, but does not have access to the identity of the relationship via w'. This is *prediction*.

4. We also require P to be able to compose two representations to create a single new representation, for which we use the notation $\boldsymbol{c} = \boldsymbol{r}_1 \circ \boldsymbol{r}_2$, where P has produced a new representation $\boldsymbol{c}$ from $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$.

5. For each of the relations in points (1), (2), and (3) above, we run P to create a composed representation, $\boldsymbol{c}$, of the query and support representations. For the known valid relationships, point (1) above, we run P to produce $\boldsymbol{c}_{t_t} = \boldsymbol{r}_x \circ \boldsymbol{r}_{y_t}$ and $\boldsymbol{c}_{t_f} = \boldsymbol{r}_x \circ \boldsymbol{r}_{y_f}$. For the known invalid relationships, point (2) above, we run P to produce $\boldsymbol{c}_{f_t} = \boldsymbol{r}_x \circ \boldsymbol{r}_{y_f}$ and $\boldsymbol{c}_{f_f} = \boldsymbol{r}_x \circ \boldsymbol{r}_{y_t}$. For the unknown relationships, point (3) above, we run P to produce $\boldsymbol{c}_{u_t} = \boldsymbol{r}_x \circ \boldsymbol{r}_{y_t}$ and $\boldsymbol{c}_{u_f} = \boldsymbol{r}_x \circ \boldsymbol{r}_{y_f}$. We seek the P that minimizes the distance between each predicted valid composed relationship with a known valid composed relationship (i.e., $\boldsymbol{c}_{u_t}$ to $\boldsymbol{c}_{t_t}$) and maximizes the distance to a known invalid composed relationship (i.e., $\boldsymbol{c}_{u_t}$ to $\boldsymbol{c}_{f_t}$), and vice-versa for the predicted invalid composed relationships. In this way, we aim for P to create an interlocking set of constraints to encourage consistency within the environment.

P is always run within each environment without direct access to the other environments[8]. We seek the P that additionally minimizes the distances between composed representations across all environments in $\mathcal{E}$, between the nearest valid relationships and invalid relationships, and maximizes the distances between mismatched valid/invalid composed representations.

With the above constraints, we then evaluate and choose the minimally sized P (i.e., most par-

[5] It is possible that some notion of *learning* and this notion of AI might turn out to be equivalent, or at least, not meaningfully distinct, but this brief blog post is too informal to pursue that further. (It is possible that the abstraction presented here amounts to seeking the most effective, most parsimonious single model across all machine learning tasks.)

[6] An equivalent, and possibly more succinct, way to say this is we seek the minimally sized (i.e., most parsimonious) program that can most effectively impute all randomly (according to some distribution) elided environments. The distinction is mostly not meaningful when talking at such a high-level as here, but here we are thinking in terms of imputation, rather than in terms of an agent in RL terms. We leave comparisons to other existing AI abstractions to a future post. However, at least as we have written it here, note that P does not directly modify the environment; some external process (not necessarily "intelligent") would be used to modify the environment constrained by the knowledge/information in P, which would then result in a new environment.

[7] There are various equivalent ways to describe such a program. This is one such way.

[8] This serves as a guard against P simply memorizing all w' across environments.

simonious, least information/bits) that produces the most compact representations with the shortest distances between representations of relationships of matching parity (i.e., valid or invalid, which we can check with access to w) and largest distances between representations of relationships of mismatched parity, across all of $\mathcal{E}$. In this way, we seek the most parsimonious model[9] that can produce the most parsimonious representations with the smallest error among known relationships, an endeavor out of which we hope new, otherwise unknown relationships become discovered, consistent with the known relationships. As we discover increasingly preferable P, according to the above, we discover an increasingly "intelligent" program that eventually reaches the point of imputing the world to the extent that it can be imputed.

In summary, in this convention, the goal of AI is to uncover new knowledge in an environment, maximizing consistency with the existing known relationships, using the smallest possible program, generating the smallest possible representations, through a process of minimizing and maximizing distances to those representations, generalizable across environments. Note that no definition of "intelligence" directly relatable to human intelligence or biology is provided, as it is not needed in the above.

That leaves *a lot* of degrees of freedom, using somewhat questionable notation. For example, there is no notion of complexity above; nor additional details accounting for the distributions of unknown, valid, and invalid relationships; nor a procedure for how exactly we would encode, or otherwise, construct the environments. More generally, this is sufficiently abstract that one's own mileage in using this to inform downstream work may vary, and certainly we might reasonably question how useful these and other AI abstractions are for concrete research.[10] However, it is not immediately clear to us why something along the lines of the above would necessarily be less useful than re-casting the goal of AI research to be recapitulating human intelligence, which itself is an under-defined topic not yet well understood by science.[11] To put it another way, since at the end of the day we are primarily only concerned with such abstractions to the extent they can serve as a guide for building real-world engineering tools, it is not a forgone conclusion that anthropomorphizing the abstract goal is the most cogent and productive path.

[9] This does not preclude the possibility that a much larger program would be used to find the smaller program.

[10] In contrast, the goal of our near-term concrete research is more simply to discover efficient learned models practically useful for particular well-defined engineering tasks in language, medicine, vision, etc. That said, with a bit of gymnastics (and a great deal of charity to the abstraction), one might see how concrete ideas from representation learning and metric learning would be encompassed by the proposed AI abstraction presented here, bound by principles of Occam's razor and resource conservation.

[11] Similarly, when particular models are criticized for accessing much more data than a human does in a lifetime, it is not immediately clear why that would be a meaningful point of comparison and criticism, unless one has already decided on the goal of emulating humans (which again, could in fact turn out to be the right goal to have). Of course, there may be other reasons to seek to minimize the amount of data given to a program that have nothing to do with human intelligence, such as reducing the total amount of resources consumed.