universe

MDPI

Article

Entropy of Artificial Intelligence

Tamás Sándor Biró * and Antal Jakovác

Wigner Research Centre for Physics, 1121 Budapest, Hungary; jakovac.antal@wigner.hu
* Correspondence: biro.tamas@wigner.hu

Abstract: We describe a model of artificial intelligence systems based on the dimension of the probability space of the input set available for recognition. In this scenario, we can understand a subset, which means that we can decide whether an object is an element of a given subset or not in an efficient way. In the machine learning (ML) process we define appropriate features, in this way shrinking the defining bit-length of classified sets during the learning process. This can also be described in the language of entropy: while natural processes tend to increase the disorder, that is, increase the entropy, learning creates order, and we expect that it decreases a properly defined entropy.

Keywords: entropy; artificial intelligence; deep learning

1. Introduction

The purpose of this article is to present a certain view of understanding and (pattern) recognition, based on the bit length of a unique but ordered coding that distinguishes between objects belonging and not belonging to a given class. Once done, we will define an entropy measuring the effectiveness of the concepts we use for classification, and show that learning is equivalent to decreasing the entropy of the representation of the input.

One of the bottlenecks in today's artificial intelligence (AI) algorithms is to conceptualize the unknown in a way which makes it automatically implementable by an AI. In this respect, the borderline recognitions after a learning process also belong to the category "unknown". A flexible AI should be able to determine when it did not recognize something and, accordingly, to run a safety protocol.

The person pushing a bicycle is neither a pedestrian nor a vehicle, still he or she should not be overrun by an automatic self-driven car. To render to one of the pre-defined classes an unexpected, so far unexperienced, and briefly unknown perception is neither smart nor intelligent. Intelligence starts where such indefinite situations are recognized and acted upon.

We model in the present article a finite universe of objects, each indexable (countable), and referred to by an at most N bits long binary digital code. These 2^N possible objects can be divided into two categories in the simplest version: 2^{N_1} belonging to a pre-defined and recognized class, and all the $2^N - 2^{N_1}$ others to the unrecognized ones. This situation is analogous to the division of phase space in statistical physics to a subsystem under observation and to an unobserved environment. Since the best separation minimizes the correlations among these two parts, the minimum in mutual information indicates the ideal subsystem—environment partition. For infinitely large systems, this leads to the canonical description, for example, by fixing the average energy in the subsystem being equal to that of the reservoir system. In finite systems, there are fluctuations around this value, and the temperature cannot be sharply defined.

The structure of this article is as follows. We start with the motivation that a comprehensive view of understanding and recognition is important for improving both natural and AI systems. Next, we turn to a mathematical formulation of understanding supported with proofs and examples. Then, we turn to the definition of the entropy of learning, proposing a formula that is minimal in the optimally trained state of the intelligent actor.

Citation: Biró, T.S.; Jakovác, A. Entropy of Artificial Intelligence. *Universe* **2022**, *8*, 53. https://doi.org/10.3390/universe8010053

Academic Editor: Lorenzo Iorio

Received: 10 December 2021
Accepted: 14 January 2022
Published: 16 January 2022

We also demonstrate how a scientific method of dealing with a few unknowns is opposed to AI methods, where complex data must be handled in repeated steps. Then, learning as a time evolution of bitwise arrangements will be pictured and related to phase space shrinking—eventually to a concept of general entropy.

What Does Science Teach Us about Learning?

The learning process, during which new knowledge is acquired, is of primary importance in human evolution and societal development. Concepts of learning, understanding and wisdom were and are anxiously debated in the history of philosophy [1]. In the natural sciences, in particular in the modern ages when these became released from the oppression of philosophy, understanding frequently appears as a simplification. Not that simplification alone would be an understanding, but the fundamental relations, newly discovered, could only be seen by paying the price of simplification. It is also an abstraction from unnecessary circumstances-but which exactly are unnecessary is more clear after one has found the right models.

In this vocabulary, knowledge is data, such as facts known about systems and processes, eventually comprised into finite bit-strings, and understanding is the model which is able to organize these facts in such a way to improve the speed and sharpness of our predictions. These predictions form a basis for interaction, transformation and, in the end, for technology. We declare an understanding when we divide the facts into two classes: relevant and irrelevant ones. Relevant for an understanding of the behavior, and relevant for being able to predict and influence future states of the piece of world under our study.

Complex systems are also often treated in science as collections of their simpler parts, with subsystems described by shorter bit-strings. When the total is nothing more than a simple sum of its ingredients, then analytic thinking triumphs. In some cases the "simple addition" may be replaced by more sophisticated composition rules, but any rule which defines a mathematical group or semi-group is associative. For associative rules, a formal logarithm can be derived which is then additive [2]. Exactly the utilization of formal logarithms defines a powerful generalization of the entropy concept: the group entropy [3].

Knowing the parts of a system and all of the interactions between those parts constitute a complete analytic model. The subsystems being simple it is only a question of computing power to make all possible predictions. Yet, we do not need them all, only the relevant ones. The goal of a Theory of Everything, the ultimate string model meanwhile was shattered by the enormous number of the possible ways it could be connected to reality, or at least to its most prominent representative, the Standard Model in particle physics [4]. With a slight extension, on the cost of phenomenological parameters, even gravity may be added to the Standard Model [5]. Accepting two dozen unexplained parameters, in principle, the Standard Model is understood and its predictions are experimentally verified. More worry arises upon not detecting deviations from it. Therefore, the Standard Model itself needs an explanation.

Computing power until the mid of 20th century was exhausted by formula writing and solving with pencils and paper. Since the dawn of computers, machines took over the bulk of computations with a speed surpassing all previous dreams. Even computer simulations of complex systems with nonlinear chaotic dynamics arose and man-made intelligent networks work by classifying complex data patterns. But one still doubts that computing machines would understand what they compute.

How to understand then complex systems? Historically, the first attempt to treat complex systems was by perturbation theory. This was doable even with paper and pencil tools. A simplified problem, solved analytically, is varied by adding small perturbing effects and the real behavior is determined by a series of further computations gradually. The main limitation of this strategy appears when perturbation series become divergent, like the quantum theory of strong interaction Quantum Chromodynamics (QCD) for processes at low momentum transfer. For such processes a massive use of computer power seemed to be the solution: as lattice gauge theory spread since the 1970s.

Such large scale simulations provide bits and numbers which have to be interpreted. These are virtual experiments not telling us more about the essentials than a real world experiment—but with a lower cost and higher insecurity about their relevance. That circumstance drives their proliferation. So why is it that knowing a bunch of numbers does not mean understanding?

An important example from high energy physics is the missing proof of the existence of a mass-gap in quantum Yang–Mills theory, declared to be one of the millennium problems by the Clay Institute, despite the numerical evidence. In further examples, the underlying level looks simpler, more understood, than the composite one: QCD looks theoretically simpler than nuclear physics, the Schrödinger equation than chemistry, the structure of amino acids than the mechanism of protein folding. Networks of simplified neurons are also more easily simulated than thinking and other higher brain functions.

A further difference lies in the following: a scientific model is usually understood term by term, all elements of basic equations represent separate physical actors. For example, in hydrodynamics, the pressure, energy density, shear and bulk viscosity all have their separate roles. In simulating complex dynamics, the individual terms are frequently just auxiliary variables without any special meaning for the behavior of the entire system. Completely different representations may yield the same result; in this way they are equally good. What did we understand here?

We would like to mention now some challenges for the AI learning. Despite the overwhelming popularity of deep neural networks (DNN) [6] and other machine learning (ML) approaches, we do not grasp why a DNN performs better than the analytic, simplifying method of science. In table games (chess, go, nine men's Morris, etc.) or by face recognition, DNNs are effective and are already faster than people, whose evolution prepared them for face recognition. What is common and what is different in image classification and solving Newton equations?

Mathematically, a feed-forward deep neural network is a series of functional mappings, between the input x and output y in the form of $y = \mathcal{N}(x, W)$, whose parameters, W, weight possible paths of signal propagation. Learning is then a re-weighting, at the end of which the output y significantly and repeatedly differs for inputs belonging to one or another class, that we wanted to teach the AI. Such algorithms mostly lead to the required result, but still their performance does not seem to be based on a simplified model: their results can be repeated by copying all weights to another AI, but lesser or erroneous copies ruin the whole procedure soon.

Such problems reveal themselves when neural networks make errors unexpectedly. The well-trained set of weights are also the most vulnerable ones to adversarial attacks. Human knowledge, once learned, seems to be more robust against such pernicious effects. This leads us to a conclusion that present day AIs recognize and interpret their environments differently.

The learned weights are not in a one-to-one relation to understanding: some networks, if shown patterns in different order or trained from different initial states (reflecting a various history of previous learnings), end up with different weights. Not even close to each other. Is then DNN learning chaotic? An entropy producing process? That is hopefully a false conclusion. Details of patterns must be unimportant for recognition. The basis for functioning well is a distinction between the relevant and irrelevant combinations inside the complex data sets.

We want to describe an example of what understanding may mean for humans and AI systems, respectively. Let us follow how a picture shown in Figure 1 exhibiting a girl and her mother appears to a computer. The digital image, describing the pixel colors in some coding, is a complete description at a given resolution, say 1 Megapixel. In the real world image of the painting there were brush strokes and chemical paints instead of pixels. That is the way how art copying works. A more economical and simpler procedure would be to name the objects seen on the canvas and their relations.

Figure 1. Pino Daeni: Summer Retreat.

The analytic method defines subsystems in the whole, naming a girl, a woman, clouds and sand, a wind blown red dress, and so forth. Repeating this analysis to smaller and smaller subsystems of subsystems one could reach beyond the one Megapixel description in data size. However, not all details are necessary for recognition and categorizing; in the human world, a depth of a few levels is sufficient. "Summer retreat", the title of the painting, comprises the relevant information in this case.

It is demonstrated by this example that the same thing can be analyzed in different ways, attaching coordinates in the space of possible objects or possible positions and colors of pixels, too. Both serve as an acceptable reconstruction of the same image. Understanding means, however, that we select common features of all images inside a given collection of them. Saying that the top left pixel is red can be less relevant than indexing it as a Pino Daeni painting with a mother and her daughter on a beach.

By knowing what is common in several images, it is much easier to decide whether a randomly drawn exemplar belongs to our cherished collection or not. Or asking for a characteristic example from the chosen collection (category) of images, a drawing with a cat passes the test while another one with a dog would not. AI systems must perform exactly such jobs.

In all of these examples, recognition is mapped to a selection of subsets in a larger set. We test the understanding by performing AI tasks: classification, regression, lossless data compression, encoding.

There are preceding works in computer science dealing with the description of understanding. The hunt for mathematizing or at least algorithmizing the cognitive abilities of humans dates back to the beginning of learning theory [7,8]: an approximately correct model is defined mathematically. The closest to our approach is representation learning [9,10] where the aim is to select an appropriate representation for complex inputs which optimally facilitates the design of a machine learning architecture. Representations can be disentangled using symmetry groups; for recent works, see [11].

There are several attempts to go beyond the limitations of present-day AI (for example and for references cf. [12]).

The physics example is the use of one exact renormalization group (RG) [13]. In computing science, RG methods were applied by [14,15] and in connection with Boltzmann Machines in [16].

2. Our Mathematical Model Space

Before referring to mathematical definitions, detailed in [17], let us construct a little example. In order to follow all steps we consider a three bit universe, containing $2^3 = 8$ possible elements. Our analogon to the total phase space in statistical physics is now this finite set,

$$\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\}. \quad (1)$$

To be more precise, we have here a coordinatization (representation)

$$x : \Omega \to B^3, \quad (2)$$

where $B = \{0, 1\}$ is a binary set. The elementary bits of an element $\omega \in \Omega$ are $x_3(\omega), x_2(\omega)$, $x_1(\omega)$, which corresponds to the "pixel-wise" representation of the set.

These 3-bit strings describe all possible cases (objects), which may occur for recognition. Now we select our "cats". Let all kinds of cats belong to the subspace

$$\Omega_1 \ = \ \{001, 010, 100, 111\}. \tag{3}$$

This example is chosen in a smart way: (i) The number of cases are half of all possible ones; this is reflected in the ratio of the respective cardinalities (sizes) $|\Omega_1|/|\Omega| = 1/2$; (ii) the number of set bits in the chosen subspace (value 1) are exactly the half of all bits, so Ω_1 is bit balanced; (iii) in this way both in the total set and in the subset, the expectation value of a uniformly randomly chosen bit is $1/2$.

Are these bits all independent? Are all elements (images) equally probable? The answer is yes for the second question, but no for the first question. It can easily be checked by looking for joint probabilities for two and three bits: whenever those factorize to a product of one-bit-probabilities, there is no correlation. The probability that is relevant here is the conditional probability for the Ω_1 subset:

$$\mathrm{Prob}_{\Omega_1}(x_1 = \sigma_1, x_2 = \sigma_2, x_3 = \sigma_3) = \frac{1}{|\Omega_1|} \sum_{\omega \in \Omega_1} \delta(x_1(\omega) = \sigma_1)\delta(x_1(\omega) = \sigma_1)\delta(x_1(\omega) = \sigma_1), \tag{4}$$

where $\delta(a = b) = 1$ if $a = b$ and 0 otherwise (indicator function or Kronecker-delta).

In our example, Ω_1 above, the two-bit probabilities factorize:

$$\mathrm{Prob}_{\Omega_1}(x_1 = 1, x_2 = 1) \ = \ \frac{1}{4} \ = \ \mathrm{Prob}_{\Omega_1}(x_1 = 1) \cdot \mathrm{Prob}_{\Omega_1}(x_2 = 1) \ = \ \frac{1}{2} \cdot \frac{1}{2}, \tag{5}$$

and similarly for any further pairs of two bits in the sample Ω_1. However, the coincidence of all three bits in this set is also $1/4$ (a single element from the four possible ones in our subset), while the product of independent probabilities would be $(1/2)^3 = 1/8$:

$$\mathrm{Prob}_{\Omega_1}(x_1 = 1, x_2 = 1, x_3 = 1) \ \neq \ \mathrm{Prob}_{\Omega_1}(x_1 = 1) \cdot \mathrm{Prob}_{\Omega_1}(x_2 = 1) \cdot \mathrm{Prob}_{\Omega_1}(x_3 = 1). \tag{6}$$

The bits in this example are pairwise independent, but not entirely independent. In pattern recognition of megapixel images it is also typical that on a few bit level they might seem independent, but not as a whole.

The complementer set, the "non-cat" images show the same property. For

$$\Omega_2 \ = \ \{000, 011, 101, 110\}, \tag{7}$$

all two-bit joint probabilities factorize, but $\mathrm{Prob}_{\Omega_2}(x_1 = 1, x_2 = 1, x_3 = 1) = 0$, since there is no element with all bits set there. Moreover, the sets Ω_1 and $\Omega_2 = \Omega \backslash \Omega_1$ are also not independent of each other: the elements in Ω_2 are the bitwise negations of the elements in Ω_1, just in a varied order.

Since the subsets Ω_1 and its complement Ω_2 are smaller than the total, they can be mapped onto shorter codes. We may choose the index number in the above listings. For the "cat" subset we obtain:

$$\{001, 010, 100, 111\} \ \rightarrow \ \{00, 01, 10, 11\}, \tag{8}$$

and for its complementary set (containing the bitwise negated elements) the same indexing applies:

$$\{000, 011, 101, 110\} \ \rightarrow \ \{00, 01, 10, 11\}. \tag{9}$$

The new coordinates, $y_i \in \{00, 01, 10, 11\}$, are coincidentally the first two-bit combinations in the original subsets. It follows that the original images differ only in their third bits, that

is, the relevant bit. On the other hand, there is no one-bit decision: the third bits are again set or unset with equal probability.

The common set with two-bit strings finally has to be supplemented by a third bit deciding whether "cat" or "non-cat." We set an extra bit in the leftmost position if "cat" and unset it otherwise. That is a bijection among the elements of the total set:

$$\Omega^{\text{new}} \;=\; f(\Omega^{\text{old}}) \tag{10}$$

occurs in a way that from the original set given in Equation (1), we obtain a re-ordered one,

$$\{non, cat, cat, non, cat, non, non, cat\} \leftrightarrow \{non, non, non, non, cat, cat, cat, cat\}. \tag{11}$$

The subset for cats with the leading bit set is given in this example:

$$\Omega_1^{\text{new}} \;=\; \{100, 101, 110, 111\} \tag{12}$$

together with its "non-cat" complement subset,

$$\Omega_2^{\text{new}} \;=\; \{000, 001, 010, 011\}. \tag{13}$$

In the newly ordered total space, Ω, now the non-cat subset is listed first and the cat subset second in the natural index order,

$$\Omega^{\text{new}} \;=\; \{000, 001, 010, 011, 100, 101, 110, 111\}. \tag{14}$$

After these replacements in the subsets, the first bit alone decides whether the element is a cat or a non-cat. At the same time, these new subsets are totally decorrelated. We have for Ω_1^{new} the following one-bit, two-bit and three-bit probabilities:

$$\begin{aligned} \text{Prob}_{\Omega_1^{\text{new}}}(x_1 = 1) = 1, \qquad \text{Prob}_{\Omega_1^{\text{new}}}(x_2 = 1) = 1/2, \qquad \text{Prob}_{\Omega_1^{\text{new}}}(x_3 = 1) = 1/2, \\ \text{Prob}_{\Omega_1^{\text{new}}}(x_1 = 1, x_2 = 1) = 1 \cdot 1/2 = 1/2, \\ \text{Prob}_{\Omega_1^{\text{new}}}(x_2 = 1, x_3 = 1) = 1/2 \cdot 1/2 = 1/4, \\ \text{Prob}_{\Omega_1^{\text{new}}}(x_1 = 1, x_3 = 1) = 1 \cdot 1/2 = 1/2, \\ \text{Prob}_{\Omega_1^{\text{new}}}(x_1 = 1, x_2 = 1, x_3 = 1) = 1 \cdot 1/2 \cdot 1/2 = 1/4, \end{aligned} \tag{15}$$

proving total independence. The same applies for the complement subset, as it is easy to verify. One concludes that the perfect learning reduced the bit-correlations to zero in both subsets. The change in the entropy of the cat subset is also reflected in the change of the random bit expectation value. In the cat subset originally, it was $\mathbb{E}(x \,|\, \Omega_1^{\text{old}}) = 1/2$, while in the learned state it is increased to $\mathbb{E}(x \,|\, \Omega_1^{\text{new}}) = 2/3$. For the complement non-cat subset, one obtains accordingly a decrease of this bit expectation value from $1/2$ to $1/3$.

Summarizing the lesson from this example, learning is equivalent with an ordering among the permutations describing all possible orders (i.e., indexing) of the objects in an AI universe. After selecting and resetting the significant bits, the number of possibilities inside the subsets (cat and non-cat in the present example) shrinks.

Now, we generalize the concept of understanding, as detailed in ref. [17]. The finite but huge embedding set Ω contains the subset Ω_1 whose elements we will identify as recognized. Rearranging the elements of the big set in a way that the subset elements are in a common block, i.e., having assigned a new position by the leading relevant bits set to the value 1, and for all of the others unset, that is, value 0, is a coordinatization. This coordinatization is bijective, hence in principle reversible and therefore entropy conserving. The mappings $\xi : \Omega \to B^N$, with B containing the alphabet 0 or 1 in the digital case and N being the minimal length of bit-strings when counting for all elements (2^N is the smallest supremum for $|\Omega|$), are coordinatizations. There are $|\Omega|! = (2^N)!$ possible coordinatizations, the number of all permutations of 2^N elements.

Accordingly, in Ref. [17], the following definitions were given:

def.: A *complete model* of a subset of all possible images, $C \subset X$, is a bijection to an N-bits string, whose coordinates (single bits) are totally independent. The $\mathrm{Prob}_C(x = \sigma)$ probabilities are either deterministic, showing the value of zero or one, or are uniformly distributed—for irrelevant bits.

Similar notions are defined for an ensemble of disjoint sets. Relevant and irrelevant bits are defined as follows:

def.: The deterministic (zero or one) bits in the 'cat' set C are *overall relevant coordinates*,
def.: while *partially relevant coordinates* are those independent bits that are either deterministic or uniformly distributed for all C_a and C, but at least in one subset, they are deterministic.
def.: *Irrelevant coordinates* are those independent coordinates that are uniform distributed in all subsets which are 'cats.'

A permutation of either the irrelevant or the relevant coordinates among themselves, maintaining probability distributions, will not change a complete model. If one provides a complete model of the subset $C \subset X$, we define that as understanding in the present framework.

In the above cited Ref. [17], it was proven that a complete model always exists, as well as a common complete model for pairwise disjoint sets.

We can also demonstrate, following Ref. [17], that knowing a complete model makes all AI tasks trivial:

- Classification: In order to find elements from disjoint subsets and put them apart, the partially relevant bits have to be inspected. Moreover, if the leading bits disagree with those of the union set, then one immediately concludes that the shown image is not an element of any pre-determined class: an outlier is identified.
- Regression: i.e., obtaining parameters of a function from noisy function values can also be treated as a classification problem. Let, e.g., the sets $\Omega(a, b)$ contain the noisy functions around the smooth one with parameters a and b. A pair (x_i, y_i) belongs to $\Omega(a, b)$ if the probability, derived by using a model of the noise, is maximal. Finally, once a common complete model is learned, one decides about any further point pair by inspecting the partially relevant bits. These also indicate if the found numerical values do not fit in any of the classes. The AI may understand when it does not understand. Will that imply intelligence or awareness?
- Decoding: the AI task is to single out a random 'cat,' a random element in Ω_i. Since the relevant bits are constant over Ω_i, one performs:

$$x^{-1}(\sigma_{relevant}, \sigma_{irrelevant} = \text{random uniform}) \in \Omega_i. \tag{16}$$

  The distribution of irrelevant coordinates being uniform, this chooses among the elements with equal probability.
- Data compression: By knowing that the relevant bits are all the same for the cats, it suffices to keep the irrelevant ones, compressing the required length of bit-strings this way. This compression is lossless and can be undone.

We note here, following Ref. [17], that model building and understanding in natural sciences also can be viewed as separating relevant and irrelevant information, compressing this way the amount of data necessary for classification. By evaluating experimental results and by planning new experiments, the researcher's goal is to disentangle relevant information from the irrelevant ones.

A serious step towards a deep understanding also occurs when correlations appear between physical quantities, assumed to be independent at first glance.

3. Entropy of a Representation

As we have discussed, learning is a process by which we try to find a coordinatization, where the coordinates are independent, and where the number of relevant coordinates is maximal. When we have understood a subset, that is, when we can provide a complete model for that, we have organized all information in the most ordered and most compact way. On the other hand, if we made mistakes with random changes in the coordination, then the effectiveness of the representation would be worsened, proving that random bit changes can lead the system out of the most ordered, learned status.

This suggests that we can associate an entropy to the representation of the system, which is *minimal* in the optimal, learned state, and which grows with random processes. In this sense we want to speak about the entropy of the cognitive system [18] rather than the entropy of the physical system containing the cognitive system (i.e., we speak about the entropy of the software and not of the hardware). As another approach, we note that, in the completely trained system, the information is stored in such a way that the redundant and synergetic information are minimal [19], and we try to define an entropy concept that successfully represents this expectation. We mention here that the information entropy can be converted into units of physical entropy, too. It has been used for the assessment of complexity in a number of natural objects (hardware) in Ref. [20].

We start from the Shannon entropy formula, which is defined for a general n bit system as

$$S_{Shannon}(x,\Omega_1) = -\sum_{\sigma_1,\ldots,\sigma_n=0}^{1} \mathrm{Prob}_{\Omega_1}(x_1=\sigma_1,\ldots,x_n=\sigma_n)\log_2 \mathrm{Prob}_{\Omega_1}(x_1=\sigma_1,\ldots,x_n=\sigma_n). \tag{17}$$

In the present case, all configurations of Ω_1 are equally probable, therefore $\mathrm{Prob}_{\Omega_1}(x_1=\sigma_1,\ldots,x_n=\sigma_n) = 1/|\Omega_1|$ if the given bit configuration is in Ω_1, and 0 otherwise. Therefore we find:

$$S_{Shannon}(x,\Omega_1) = \log_2 |\Omega_1| = N_1. \tag{18}$$

This entropy depends only on the number of elements in the subset, independent of the representation. Indeed, this is the goal of this definition: it represents the minimal information needed to describe the Ω_1 subset.

The joint probabilities usually do not factorize: there is a correlation between the individual bits, as we have seen earlier. In general, we expect that the bitwise distribution functions overestimate the information content of the system; by neglecting the nontrivial correlation between them. Thus, we may conjecture the inequality:

$$S_{Shannon}(x,\Omega_1) \leq S_{indep}(x,\Omega_1), \tag{19}$$

where S_{indep} is the form obtained by substituting $\mathrm{Prob}_{\Omega_1}(x_1=\sigma_1,\ldots,x_n=\sigma_n) = \prod_{i=1}^{n} \mathrm{Prob}_{\Omega_1}(x_i=\sigma_i)$; after simplifications we find:

$$S_{indep}(x,\Omega_1) = -\sum_{i=1}^{n}\sum_{\sigma=0}^{1} \mathrm{Prob}_{\Omega_1}(x_i=\sigma)\log_2 \mathrm{Prob}_{\Omega_1}(x_i=\sigma). \tag{20}$$

This formula already depends on the representation. From an information theory point of view, its content suggests that the sum of information coming from individual bits is larger than the minimal information that describes the studied subset, if the information carried by the bits are not independent. The equality in the relation (19) occurs exactly in the case if in the given representation, the individual bit information is independent, then the joint probabilities indeed factorize.

This means that, in the complete model, where only the individual bits counting for differences inside the subset are independent, the representation entropy is minimal (maximizing only over the least subset). This leads us to formulate the main proposal of this paper:

Proposal 1. *The entropy characterizing the representation* $x : \Omega \to B^N$ *over the subset* $\Omega_1 \subset \Omega$ *is*

$$S_{repr}(x, \Omega_1) = -\sum_{i=1}^{n} [p_i \log_2 p_i + (1 - p_i) \log_2 (1 - p_i)] \tag{21}$$

where $p_i = \mathrm{Prob}_{\Omega_1}(x_i = 1)$ *and therefore* $1 - p_i = \mathrm{Prob}_{\Omega_1}(x_i = 0)$. *The minimal value of* S_{repr} *is the Shannon entropy* $\log_2 |\Omega_1|$. *A process that decreases the representation entropy leads us towards a well learned state (to a complete model).*

Let us demonstrate how this works in the case of our three-bit example. Here $|\Omega| = 8$ and $|\Omega_1| = 4$, thus the value of the Shannon entropy is 2. We already calculated the one-bit probabilities in the complete model in (15): $p_1 = 1/2$, $p_2 = 1/2$ and $p_3 = 1$. The representation entropy of the complete model,

$$S_{repr}(x, \Omega_1) = -\sum_{i=1}^{2} \left[\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right] + [1 \log_2 1 + 0 \log_2 0] \; = \; 2, \tag{22}$$

is therefore indeed equal to the Shannon entropy. The representation entropy of the original choice, cf. Equation (3), with $p_1 = p_2 = p_3 = 1/2$ is:

$$S = -\sum_{i=1}^{3} \left[\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right] \; = \; 3, \tag{23}$$

which is larger than the Shannon entropy. In this example, altogether we have $\binom{8}{4} = 70$ representations, thus we can calculate the representation entropy for all of them, and verify that indeed only the independent cases possess minimal entropy. There are 6 such cases: the relevant bit can be the 1st, 2nd or 3rd, and its value can be 0 or 1.

In general, if we have N bits in total, and N_1 bits for representing Ω_1, the number of all representations is given by $\binom{2^N}{2^{N_1}}$. Among these vastly large number of representations the number of the complete models is $\binom{N}{N_1} 2^{N-N_1}$, because we have to choose the $N - N_1$ relevant bits in $\binom{N}{N-N_1} = \binom{N}{N_1}$ ways, and each relevant bit can be either 0 or 1, yielding the factor 2^{N-N_1}.

Practically, a representation is manifested as a vector of weights in a neural network. We can measure the representation entropy by approximating the bitwise distributions on a trial set. Then we can compare two representations, and we can move towards the minimum. Note that in this process there is no reference to what the image describes.

This leads to a general unsupervised learning strategy: we show images to the AI, then, by minimizing the representation entropy, it will be able to tell what the common features of the shown set are by setting the relevant bits. For example, to train a self-driving car we shall equip a normal car with a camera that records the images that belong to the "normal" view. Then, as the AI learns what a "normal" condition means, it will recognize if something is "abnormal." The advantage of this process is that we do not have to tell *what* can cause a misfunction, because the AI learns what "normal" means, and automatically classifies any other images as potential danger.

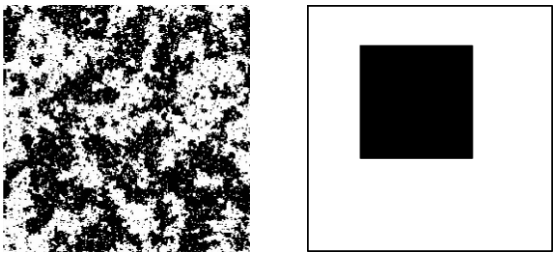
4. How Many Relevant Bits?

Scientific models have typically just a few relevant bits; psychologists claim that a person can keep at most seven different concepts simultaneously in mind. In contrast to that, the typical AI learning, recognizing and classifying task meets a huge number of both relevant and irrelevant bits.

We may even compare images and science models, by presenting an input set of 0 or 1 pixels.

In the Ising model, binary states are fixed along a chain or array and pairwise interactions are assumed, c.f. Figure 2. In statistical physics modeling, one averages over a large

scale with Boltzmann weights, $e^{-\beta H}$, with a parameter β and a Hamiltonian, H, describing the interaction energy of aligned and misaligned spins of elementary magnets. Only the relevant quantities occur in this modeling process: the value of the parameter β, the number of misaligned pairs (01 and 10 bit pairs), and the total number of set bits, describing the overall magnetization. Setting white for 0 and black for 1, one produces the typical image.



(**a**) (**b**)

Figure 2. One bit pixel images obtained from (**a**) magnetic Ising model fluctuations, and (**b**) a black square on white background.

For the 'artistic' picture of a single, arranged black square, we may have the same number of set and unset bits, the overall recognition, however, differs a lot. When an AI has to recognize and single out the square, which in the Ising model is a very improbable, yet not impossible configuration, then a number of irrelevant bits will be studied first.

For compressing the data of the square, only a few bits of information are needed, e.g., the coordinates of a corner and the side lengths. The bits needed for compression are the irrelevant bits. For the description of this image, the pixel information is not too informative, since the image is very correlated. So, while for the Ising model the color of each pixel is relevant, in case of the black square, a serious re-coordinatization is needed: instead of pixels, we speak about a "square," a nonlocal object from the point of view of the pixels.

Either the relevant or the irrelevant coordinates are few; one has an enormous compression possibility and with that a good chance for understanding. That is the basis for the hope that by applying AI methods, most importantly image recognition algorithms, we may learn new information about physical (and chemical, biological, ecological, medical, economical, social) problems of high complexity.

In typical image recognition tasks, both the relevant and irrelevant coordinates are numerous; therefore, only their collective effect carries distinctive meanings. Various deep neural networks with vastly differing weight factors can have similar performance on a given image set.

In natural learning, to start with, the complete subset which has to be learned is unknown. In smaller sets almost all parameters (describing bits) can be relevant. Later, getting inputed with a large number of examples, the set growth, and the number of irrelevant bits grow with this process too. For recognizing N irrelevant bits, we have to see about 2^N images.

The time evolution of the bits during the learning process includes the variation of irrelevant bits. Their versatility describes the speed of learning. The never changing coordinates do not need to be remembered: all are the same all of the time. Biological learning is very probably accompanied by a lossy compression. Since a lossy compression is irreversible, the entropy must change.

In supervised learning, the relevant coordinates are told explicitly, either by laws or one by one. That accelerates the learning, but it will not change the nature of the lossy compression. Thus, a large number of unannotated inputs has to be applied and trained according to the measure of relevance, and to the time variability of a given coordinate. This way we do not transplant our own knowledge to the AI; it will build its own system of notions and understanding of its inputs.

Some bits may seem relevant while they are not, e.g., due to an unfortunate sampling that underrepresents or overrepresents features. Superstition may work this way, and the future AI reaching its performances in understanding its input world comparable to ours, might become superstitious.

Another possibility of a false understanding occurs when one omits a bit as irrelevant, which is in fact relevant. Deterministic chaos may appear as a genuine underlying randomness in the universe. This randomness seems unexplorable and unpredictable, as frequently assumed by miscalulations of the financial markets. Identifying this error and promoting the corresponding coordinate to a relevant one in our descriptive model is the "aha" effect of sudden understanding.

5. Conclusions

In this article we presented a bitwise picture of understanding and classification. The AI tasks were viewed based on categorization as identifying relevant and irrelevant bits in bitstring codes of set elements, most prominently images. Relevant bits have a deterministic value, 1 for the selected set and 0 for the non-set. Irrelevant bits have a random value with a uniform distribution of ones and zeros.

The result of learning is a coordinatization, a re-arrangement of bits where the relevant ones constitute a block. This can be described as a bijection from the original indexing (original showing order of training images) and therefore does not change the original Shannon entropy.

However, we propose considering another quantity, the representation entropy, whose construction assumes bitwise independence among irrelevant bits. Its value coincides with the Shannon entropy only in a completely learned state, when the probability of all bits in fact factorizes. The actual representation entropy is larger before learning and reaches its minimal value after a complete understanding. This quantity may serve as a basis of algorithmic strategies for unsupervised learning; its decrease would gradually lead to a state of decorrelated single bits, providing, in this way, a maximal compression of code length to remember. The most intriguing novel element in this strategy is to teach AI systems to recognize that they do not recognize certain images. This can then be a basis for automatic switching to safety protocols, like stopping self-driving cars.

The relevant–irrelevant coordinatization is not unique. The bitwise independence in the learned state only restricts it. Whenever the number of either relevant or irrelevant bits are just a few, humans can build models of understanding easily. Having a huge number of both types of coordinates on the other hand cries out for using AI algorithms to disentangle cats form non-cats.

Finally, we note that our concept about ML as a reduction of the number of random bits by identifying a subset of pictures may show some analogy to classical compositional data analysis [21]. However, the entropy reduction associated with the learning process is a stand-alone concept, not requiring the complexity and intricacies of that mathematical field.

Author Contributions: Both authors contributed equally to concept development, calculations, discussions and manuscript preparation. T.S.B. raised some funding. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Hungarian National Bureau for Research, Development and Innovation NKFIH (OTKA) under the contract No. K123815.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors acknowledge useful discussions with D. Nagy, A. Telcs and G Orbán.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zalta, E.N. (Ed.) Steup, Matthias and Ram Neta, Epistemology. The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). Available online: https://plato.stanford.edu/archives/fall2020/entries/epistemology (accessed on 12 January 2022).
2. Biró, T.S. Thermodynamics of composition rules. *J. Phys. G* **2010**, *37*, 094027. [CrossRef]
3. Tempesta, P. Formal grouops and Z-entropies. *Proc. Math. Phys. Eng. Sci.* **2016**, *472*, 20160143.
4. Wikipedia Article. Available online: https://en.wikipedia.org/wiki/Standard_Model (accessed on 12 January 2022).
5. Shaposhnikov, M.; Wetterich, C. Asymptotic safety of gravity and the Higgs boson mass. *Phys. Lett. B* **2010**, *683*, 196–200. [CrossRef]
6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
7. Available online: https://en.wikipedia.org/wiki/Computational_learning_theory (accessed on 12 January 2022).
8. Osherson, D.N.; Stob, M.; Weinstein, S. *Systems That Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*; MIT: Cambridge, MA, USA, 1990.
9. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
10. Higgins, I.; Sonnerat, N.; Matthey, L.; Pal, A.; Burgess, C.P.; Bosnjak, M.; Shanahan, M.; Botvinick, M.; Hassabis, D.; Lerchner, A. SCAN: Learning Hierarchical Compositional Visual Concepts. *arXiv* **2017**, arXiv:1707.03389.
11. Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; Lerchner, A. Towards a Definition of Disentangled Representations. *arXiv* **2018**, arXiv:1812.02230.
12. Lu, H.; Li, Y.; Chen, M.; Kim, H.; Serikawa, S. Brain Intelligence: Go Beyond Artificial Intelligence. *Mob. Netw. Appl.* **2018**, *23*, 368–375. [CrossRef]
13. Available online: https://en.wikipedia.org/wiki/Renormalization_group (accessed on 12 January 2022).
14. Mehta, P.; Schwab, D.J. An exact mapping between the Variational Renormalization Group and Deep Learning. *arXiv* **2014**, arXiv:1410.3831.
15. Lin, H.; Tegmark, M. Why does deep and cheap learning work so well? *J. Stat. Phys.* **2017**, *168*, 1223–1247. [CrossRef]
16. Available online: https://en.wikipedia.org/wiki/Boltzmann_machine (accessed on 12 January 2022).
17. Jakovac, A.; Berenyi, D.; Posfay, P. Understanding understanding: A renormalization group inspired model of (artificial) intelligence. *arXiv* **2020**, arXiv:2010.13482.
18. Chen, M.; Hao, Y.; Gharavi, H.; Leung, V.C. Cognitive information measurements: A new perspective. *Inf. Sci.* **2019**, *505*, 487–497. [CrossRef] [PubMed]
19. Mediano, P.A.; Rosas, F.E.; Luppi, A.I.; Jensen, H.J.; Seth, A.K.; Barrett, A.B.; Carhart-Harris, R.L.; Bor, D. Greater than the parts: A review of the information decomposition approach to causal emergence. *arXiv* **2021**, arXiv:2111.06518.
20. Csernai, L.P.; Spinnangr, S.F.; Velle, S. Quantitaive assesment of increasing complexity. *Phys. A* **2017**, *473*, 363–376. [CrossRef]
21. Aitchinson, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. B* **1982**, *44*, 139–160.