Bohr's Observation about Deep Truths, Quantum Computing, Multiple-Valued Logic, Neural Networks, LLMs, Hegel's Negation of Negation (and Maybe Even Laws of Sexual Attraction)

Miroslav Svítek[1], Olga Koshevela[2],
Vladik Kreinovich[3], and Nguyen Hoang Phuong[4]

[1]Faculty of Transportation Sciences, Czech Technical University in Prague, Konviktska 20
CZ-110 00 Prague 1, Czech Republic, svitek@fd.cvut.cz
[2,3]Departments of [2]Teacherc Education and [3]Computer Science,
University of Texas at El Paso, 500 W. University,
El Paso, TX 79968, USA, olgak@utep.edu, vladik@utep.edu
[4]Artificial Intelligence Division, Information Technology Faculty,
Thang Long University Nghiem Xuan Yem Road, Hoang Mai District,
Hanoi, Vietnam, nhphuong2008@gmail.com

1. This phenomenon was first observed in physics

- One of the problems with which the 19th century physics struggled was the problem of light.
- The problem was that in some experiments, light behaves like a wave, while in other experiments, it behaves like particles.
- Both ideas led to useful conclusions, but they seemed to be inconsistent with each other.
- Either light is particles or it is not made of particles.
- Quantum physics solved this problem by providing a way to describe such phenomena in a consistent way.
- In this situation:
  - both the idea that light is a continuous wave and this idea's negation – that light is formed by discrete particles
  - turned out to be deep ideas, of great importance to physics.

2. Niels Bohr noticed that this phenomenon is ubiquitous

- Niels Bohr observed that a similar phenomenon occurs in many other areas of knowledge and belief.
- For example, in physics, we have:
  - the original idea that our space has Euclidean geometry and
  - the opposite idea – that eventually led to general relativity – that the geometry of our space is non-Euclidean.
- They are both deep and useful ideas.
- As a more general example: both the idea that there is an almighty God and the idea that there is no God.
- Both are deep and productive ideas.
- E.g., the idea that there is no God led many scientists to successfully search for natural explanation of real-world phenomena.

3. Niels Bohr noticed that this phenomenon is ubiquitous (cont-d)

- These explanations greatly contributed to modern physics, biology, etc.
- Niels Bohr summarized this phenomena by saying that "It is the hallmark of any deep truth that its negation is also a deep truth."

4. But how can we explain this phenomenon?

- Scientists and philosophers came up with many examples confirming that this phenomenon is indeed ubiquitous.
- But why? How can we explain this ubiquity?
- In this talk, we provide an explanation for the deep truths phenomenon.
- This explanation is based on two facts:
  - that in symmetry situations, under reasonable conditions, optimum is always attained in a symmetric case, and
  - that in general, maximum is attained on the border of the set of possible alternatives.

5. What do we mean by symmetries and why they are important

- The usual way we can predict future events is by using previous experience.
- We have observed that the sun rises every morning, so we predict that in a similar situation, it will rise again.
- Suppose that in the past, headache and small fever eventually led to a flu.
- Then next time, in a similar situation, the person expects that similar symptoms will lead to a flu.
- If several experiments confirmed relativity theory, then we expect the future experiments to also be in line with its predictions.
- In all these cases, we rely on the notion of similarity.

6. What do we mean by symmetries and why they are important (cont-d)

- The next time when we feel headache and small fever:
  - we may have moved to a new location,
  - we may be turning the other way,
  - we may be dressed differently.
- However, we have learned, from experience, that all these differences do not change the outcome.
- In physics, such transformations that do not affect the phenomenon that we are studying are called *symmetries*.
- Not surprisingly, symmetries are one of the main ideas in physics.

7. What do we mean by symmetries and why they are important (cont-d)

- Starting with the quark theory in the 1960s, most new theories in physics appear:
  - not in the form of differential equations – as in the time on Newton;
  - they appear in terms of the corresponding symmetries.
- Moreover, it turned out that:
  - for most previously proposed fundamental physical theories, be it electrodynamics, quantum physics, etc.,
  - their differential equations can be uniquely determined by the corresponding symmetries.

8. Towards a precise description of symmetries

- Let us denote by S the set of possible situations.
- A *symmetry* $g \in G$ is a function $g : S \mapsto S$ that maps each state $s \in S$ into a new state $g(s)$.
- Examples can be shift in space, rotation, changing clothes, changing positive electric charges to negative ones and vice versa, etc.
- Suppose that we have two symmetries g_1 and g_2.
- This means that the studied phenomenon does not change when we apply each of these transformations to the state.
- Thus, if we first apply g_1 and get $s' = g_1(s)$, and then apply g_2 and get $g_2(s') = g_2(g_1(s))$, then also the phenomenon will not change.
- This means that the composition $(g_2 \circ g_1)(s) \stackrel{\text{def}}{=} g_2(g_1(s))$ is also a symmetry.
- Similarly, the inverse transformation – e.g., if we shift back – does not change the phenomenon.

9. Towards a precise description of symmetries (cont-d)

- Thus, the class G of all the symmetries is closed under composition and under taking the inverse.
- Such classes are known as *transformation groups*.

10. What is optimization

- In its most general form, optimization means that we have some way to compare two alternatives a and b and conclude:
  - either that a is better (we will denote it by $a > b$),
  - or that b is better ($b > a$),
  - or that a and b are of equal quality with respect to the desired objective (we will denote it by $a \sim b$).
- Of course, these comparisons must be consistent.
- E.g., if a is better than b and b is better than c, then a should be better than c.
- In these terms, an alternative a_{opt} is *optimal* if for every alternative a, we have either $a_{\text{opt}} > a$ or $a_{\text{opt}} \sim a$.
- Sometimes, several alternatives are the best according to the given optimality criterion.

11. What is optimization (cont-d)

- In practice, this means that we can use this non-uniqueness to optimize something else.
- For example, a poor student may want to minimize the cost of his/her lunch.
- Suppose that there are several equally cheap options.
- This means that the student can select, among them, the one which is the healthiest (or, alternatively, the one which is the tastiest).
- This means, in effect, that the previous optimality criterion was not final.
- Now we have a new criterion in which $a > b$ if:
  - either a is cheaper than b,
  - or a and b lunches cost the same, but lunch option a is healthier.

12. What is optimization (cont-d)

- Similarly, in general:
  - if there are several optimal alternatives,
  - this means that the optimality criterion is not final, it can be detailed some more.
- Thus, for a final optimality criterion, there is exactly one optimal alternative.
- Let us describe all this in precise terms.

13. Definitions

- Let the set S be given. Its elements will be called *alternatives*.
- By an *optimality criterion*, we mean a pair of binary relations $(>, \sim)$ on the set S for which the following conditions hold for all $a, b, c \in S$:
  - if $a > b$ and $b > c$, then $a > c$;
  - if $a > b$ and $b \sim c$, then $a > c$;
  - if $a \sim b$ and $b > c$, then $a > c$;
  - if $a \sim b$ and $b \sim c$, then $a \sim c$;
  - if $a \sim b$, then $b \sim a$;
  - if $a > b$, then $a \not\sim b$;
  - always $a \sim a$.
- Let an optimality criterion $(>, \sim)$ on a set S be given.
- An alternative a_{opt} is called *optimal* if for every $a \in S$, we have either $a_{\text{opt}} > a$ or $a_{\text{opt}} \sim a$.

14. Definitions (cont-d)

- We say that the optimality criterion is *final* if there exists exactly one optimal alternative.
- Let G be a transformation group on the set S.
- We say that the optimality criterion is G-*invariant* if:
  - for every two alternatives a and b and for every symmetry $g \in G$,
  - the following two conditions are satisfied: if $a > b$, then $g(a) > g(b)$; and if $a \sim b$, then $g(a) \sim g(b)$.

15. Proposition

- *For every G-invariant optimality criterion, the optimal alternative a_{opt} is itself G-invariant, i.e., $g(a_{\text{opt}}) = a_{\text{opt}}$.*
- **Proof.** By definition of an optimal alternative, for every $a \in S$, we have either $a_{\text{opt}} > a$ or $a_{\text{opt}} \sim a$.
- In particular, for every a and for every $g \in G$, we have either $a_{\text{opt}} > g^{-1}(a)$ or $a_{\text{opt}} \sim g^{-1}(a)$.
- Since the optimality criterion is G-invariant, this implies that either $g(a_{\text{opt}}) > g(g^{-1}(a)) = a$ or $g(a_{\text{opt}}) \sim g(g^{-1}(a)) = a$.
- In other words, for every $a \in S$, we have either $g(a_{\text{opt}}) > a$ or $g(a_{\text{opt}}) \sim a$.
- By definition of an optimal alternative, this means that the alternative $g(a_{\text{opt}})$ is also optimal.
- But since the optimality criterion is final, this means that there is only one optimal alternative, so $g(a_{\text{opt}}) = a_{\text{opt}}$.

16. Optimum is usually on the border: general idea

- According to calculus, the maximum of a function on a bounded domain is attained:
  - either inside this domain – in which case this is a stationary point, i.e., a point at which all partial derivatives are 0,
  - or at the border of the domain.
- In general, a typical objective function has few stationary points.
- If the conditions describing the domain are significant:
  - this means that the domain is relatively small
  - in comparison with the set of all possible alternatives.
- When the domain is small, the probability that this domain happens to include one of the few stationary points is also small.

17. Optimum is usually on the border: general idea (cont-d)

- Thus, in the vast majority of cases, the domain does not have any stationary points inside.
- So, the maximum of an objective function is attained on the border of the domain.

18. Examples

- A natural example is a nuclear reactor.
- To function effectively, it has to have the neuron reproduction rate equal to 1:
  - if it smaller than 1, then the reaction will stop;
  - if it is larger than 1, the flow will exponentially increase and we will have an explosion.
- In this case, the optimal solution is exactly on the border of the set of safe regimes.
- This may sound like an extreme example, but more mundane examples follow the same pattern.
- For example, when people walk, they:
  - place their body in an unstable position, at the edge of falling,
  - and then put the foot down to stabilize the situation.

19. Now let us go back to Bohr's phenomenon

- To apply the above two ideas to Bohr's observation, let us describe this observation in precise terms.
- For this purpose, we need to make an important distinction between two types of statements.
- On the precise side, we have precise empirically observable statements about the physical world, statements that are either true or false.
- On the other hand, we have statements which are more philosophical, meta-statements that are not necessarily true or false.
- Examples: whether light is a particle or whether there is a God.
- Let us consider both types of statements.

20. Case of precise statements

- Let us first consider precise statements.
- Based on each true statement s, we can make predictions that will eventually turn out of be true.
- Based on its negation $\neg s$, we can also make true predictions a, since we can have $\neg s \to a$.
- However, according to logic, the only way the formula $\neg s \to a$ can be true when $\neg s$ is false is when a is true.
- In this case, we can get a without using $\neg s$.
- So, if the statement s is true, then we may have predictions a based on the implication $s \to a$.
- These predictions are not possible without using s.
- On the other hand, suppose that the statement $\neg s$ is false.
- Then there are no true statements a that we can derive from $\neg s$ that we cannot derive without it.

21. Case of precise statements (cont-d)

- Let us describe this in numerical terms.
- Let us consider all the empirically true statements that we can derive either by assuming s or by assuming $\neg s$.
- Let p be the proportion of these statements that can be derived from s.
- Then, correspondingly, $1-p$ the proporion of the statements that can be derived from $\neg s$.
- Then, if the statement s is true, then all derived new empirical statements come from s, so $p = 1$.
- On the other hand, if the statement s is false, this means that its negation $\neg s$ is true.
- So, all derived new empirical statements come from $\neg s$.
- Thus, in this case, $p = 0$.

22. Case of meta-statements

- For meta-statements s, the situation is different.
- As we have mentioned on the example of wave-particle discussions, we can make useful deduction both from s and from its negation.
- So, for meta-statements it is possible to have the proportion p to be strictly between 0 and 1.
- Now, we are ready to deal with Bohr's question: which meta-statements are the deepest?
- To answer this question, we will apply the above two approaches.
- Note that we did not formalize what exactly "deep" means, we will be able to provide our explanation without specifying this.

23. Let us first apply the symmetry approach, and show that it explains Bohr's observation

- A priori, we do not have any reason to prefer a statement s or its negation $\neg s$.
- For example, in mathematics, a hypothesis sometimes turns out to be true, and sometimes, turns out to be false.
- We may as well consider the negation $s' \stackrel{\text{def}}{=} \neg s$ as the main statement and $s = \neg s'$ as its negation.
- If we make this change, then the new value of the proportion p' will correspond to $1 - p$: $p' = 1 - p$.
- In other words, the situation is invariant with respect to the transformation $p \mapsto 1 - p$.
- It is reasonable to assume that criteria like deepness should not change if we simply swap s and $\neg s$.

24. Let us first apply the symmetry approach, and show that it explains Bohr's observation (cont-d)

- In other words, the corresponding optimality criterion should be invariant with respect to the transformation $p \mapsto 1 - p$.
- So, in accordance with the above result, the deepest statements are invariant with respect to this transformation.
- So, for such statements, we should have $p = 1 - p$ and thus, $p = 0.5$.
- In this case, we have exactly the same number of conclusions that we can make based on the statement s as we can make based on its negation $\neg s$.
- In other words, a statement is deep if and only if its negation is deep – which is exactly what Bohr's statement implies.

25. An alternative explanation based on the second approach

- Different statements s, with different proportions p, can be described by the distance from p to one of the non-meta values 0 and 1.
- A natural way to describe the distance between the two numbers a and b by $|a - b|$.
- So the distance from p to one of the values 0 and 1 can be described as $d = \min(|p - 0|, |1 - p|) = \min(p, 1 - p)$.
- As p changes from 0 to 1, this distance changes first from 0 to 0.5 (when $p = 0.5$) and then from 0.5 back to 0.
- According to the general idea of the second approach, the optimum is attained at the border of the domain of possible distances.
- In this case, it us attained where the distance is either 0 or 0.5.
- When distance is 0, we have the values $p = 0$ or $p = 1$, i.e., in effect, we have the usual non-deep statements.

26. An alternative explanation based on the second approach (cont-d)

- Thus, if we are interested in deep statements, we should take the value $d = 0.5$ – which corresponds exactly to $p = 0.5$.
- So, this approach also explains Bohr's observations.

27. First additional example: explaining Walsh-Hadamard transformations

- It is known that using quantum phenomena, we can potentially speed up many computations.
- This is known as *quantum computing*.
- E.g., Grover's algorithm enables us to find an element with a desired property in an n-element array in time $\sim \sqrt{n}$.
- However, for any non-quantum algorithm, we need in some cases n steps, and n is much larger than $\sqrt{n}$.
- One of the main ideas behind quantum computing is that we use so-called *superpositions* of different states.
- E.g., we can have superpositions of 0 and 1, i.e., stated of the type $c_0 \cdot |0\rangle + c_1 \cdot |1\rangle$, where $|c_0|^2 + |c_1|^2 = 1$.
- In general, the coefficients c_i can be complex, but in quantum computing, usually, real values are used.

28. First additional example: explaining Walsh-Hadamard transformations (cont-d)

- A natural question is: which superposition will lead to the most effective algorithm?
- In computations, there is a natural symmetry between 0s and 1s.
- It all depends on which of the two states of a bit we call 0 and which state we call 1.
- So, we have a natural symmetry here: the swapping of 0 and 1.
- From the viewpoint of this symmetry, we can therefore conclude that the most effective state would be:
  - the state that is itself invariant with respect to this transformation,
  - i.e., for which $c_0 = c_1$.

29. First additional example: explaining Walsh-Hadamard transformations (cont-d)

- Substituting $c_0 = c_1$ into the condition $|c_0|^2 + |c_1|^2 = 1$ that describes a general quantum state, we conclude that

$$c_0 = c_1 = \sqrt{2}/2 \text{ or } c_0 = c_1 = -\sqrt{2}/2.$$

- In quantum physics, states that differ by a factor whose absolute value is equal to 1 are identical.
- In effect, a physical state is an equivalence class based on this.
- From this viewpoint, the two above cases describe the exact same state

$$\frac{\sqrt{2}}{2} \cdot |0\rangle + \frac{\sqrt{2}}{2} \cdot |1\rangle.$$

- Also, for this same reason, states 0 and 1 can be represented not only as $|0\rangle$ and $|1\rangle$, but also as $-|0\rangle$ and $-|1\rangle$.
- So we can consider a more general superposition $\pm c_0 \cdot |0\rangle \pm c_1 \cdot |1\rangle$.

30. First additional example: explaining Walsh-Hadamard transformations (cont-d)

- When both signs are negative, we get the exact same state as before.
- However, when the signs are different, we get a new state

$$\frac{\sqrt{2}}{2} \cdot |0\rangle - \frac{\sqrt{2}}{2} \cdot |1\rangle.$$

- These two states form what is called *Walsh-Hadamard* transformation of the states 0 and 1.
- They are indeed the basis of most effective quantum algorithms – many of which have been proven to be optimal.

31. Applications to neural networks: explaining why initial weights are randomly chosen

- A general neural network does not depend on how we number the neurons in each layer.
- In other words, the situation is invariant with respect to any permutation of neurons in each layer.
- So, according to our general symmetry approach:
  - the initial values of the weights should be invariant with respect to all these permutations,
  - and thus, be equal.
- However, this does not make sense: due to this symmetry:
  - if the initial values of all the weights are equal,
  - all the following values will be equal too.
- So all the neurons in each layer will simply meaninglessly duplicate each other.

32. Applications to neural networks: explaining why initial weights are randomly chosen (cont-d)

- A natural solution is to use random initial weights.
- In this case, symmetry simply means that we have the same probability distribution for each weight.
- However, the actual random values will be, of course, different.
- This is exactly what is done when we train a neural network.
- In principle, we can avoid equality if we simply deterministically assign different weights to different connections.
- However, this would not be invariant.
- So, our symmetry approach explains why it it turned out to be more effective to use randomized initial weights and not deterministic ones.

33. Applications to LLMs: they need to learn to say "I don't know"

- Current LLMs are taught to always answer "yes" or "no".
- As a result, even when they are not sure about an answer, they produce a definite answer.
- And this answer is, in some cases, very different from the truth.
- This irritating phenomenon is known as *hallucinations*.
- A natural idea is to add some intermediate truth value, in additional to 0 and 1.
- Same arguments as in the previous section lead to the selection of the value 0.5 – which is both:
  - invariant with respect to swap $0 \leftrightarrow 1$ and
  - which is the farthest away from 0 and 1.
- So, a natural idea is to teach LLMs to use a 3-valued logic, with truth values 0, 0.5, and 1.

34. Applications to LLMs: they need to learn to say "I don't know" (cont-d)

- This 3-valued logic is a particular case of a more general multiple-value logic in which truth values form the whole interval $[0, 1]$.
- It is also known as *fuzzy logic*.
- To be more precise, it is the simplest case of a more general construction of fuzzy logic.
- In general, we want to supplement 0 and 1 with some other values from the interval $[0, 1]$ to form a larger set V than $\{0, 1\}$.
- Which other values should we select?
- It makes sense to select a closed set S – because:
  - if we have values $v_n \in V$ that tends to some value v,
  - then, no matter how accurately we implement all this, for a sufficiently large n, v will be undistinguishable from v_n.

35. Applications to LLMs: they need to learn to say "I don't know" (cont-d)

- Thus, even if we do not use v, we will never be able to determine that.
- So, without changing anything, we can add all limit points to the set V and make V a closed set.
- Similarly to the previous section, we can conclude that we want to select:
  - among all the sets V for which $\{0, 1\} \subseteq V \subseteq [0, 1]$,
  - the set whose distance $d_H(V, \{0, 1\})$ from the classical 2-valued set $\{0, 1\}$ is the largest possible.

36. Applications to LLMs: they need to learn to say "I don't know" (cont-d)

- Here, $d_H(S, S')$ denotes the usual Hausdorff distance between the two sets: the smallest $\varepsilon > 0$ for which:
  - all the points from S are located in the ε-vicinity of S', and,
  - vice versa, all the points from S' are located in the ε-vicinity of S.
- Such "farthest-from-$\{0, 1\}$" sets are described by the following proposition.

37. Proposition

- *For every closed set* V *with* $\{0,1\} \subseteq V \subseteq [0,1]$*:*
  - *we have* $d_H(V, \{0,1\}) \leq 0.5$*, and*
  - *we have* $d_H(V, \{0,1\}) = 0.5$ *if and only if* $0.5 \in V$*.*
- **Proof:** see paper.

38. Discussion

- In general, the same argument shows that:
  - when we are not happy with a certain set of options,
  - a natural idea is to add a new option whose distance from the set of the previous options is the largest possible.
- We start with the first option.
- Then we select the second option which is as far as possible from the first one.
- This can be intuitively considered as the *negation* of the first option.
- The third option is as far away from both first and second options as possible.
- This can be viewed the *negation of negation*.
- This is probably an explanation of the imprecise "negation of negation" idea by the famous German philosopher Hegel.

39. A speculative application: explaining Weininger's law of sexual attraction (and the ubiquity of monogamy)

- According to Weininger's simplified model of sexual attraction:
  - each person's psychological type
  - can be described as a combination of prototypical male (m) and prototypical female (f) psychologies,
  - to be more precise, as $c_m \cdot m + c_f \cdot f$, where $c_m + c_f = 1$.
- According to this model, the greatest attraction is:
  - between the people (c_m, c_f) and (c'_m, c'_f) with opposite features,
  - i.e., features for which $c_m = c'_f$ and $c_f = c'_m$.
- Let us show that this is perfectly in line with the general symmetry idea.
- Namely, in the first approximation, we can view male and female prototypes to be kind of similar.
- So we have a symmetry between m and f.

40. A speculative application: explaining Weininger's law of sexual attraction (cont-d)

- In this case, according to our result, the optimal pair of persons should be invariant with respect to this transformation.
- So, unless we have a person with $c_m = c_f = 0.5$, the only way for the pair to be invariant is when:
  - the psychological characteristics of the second person
  - can be obtained from the characteristics of the first one by swapping m and f.
- This is exactly Weininger's equality.
- The symmetry idea also explains why monogamy is ubiquitous and menage a trois is rare.
- Indeed, it is not possible to have a configuration of three people that is invariant with respect to the swap $m \leftrightarrow w$.

41. A speculative application: explaining Weininger's law of sexual attraction (cont-d)

- We can have configurations of 4, 6, etc., people which are thus invariant.
- However, in this case, they simply consist of several invariant pairs, and there is no relation between persons from different pairs.

42. Acknowledgments

This work was supported in part by:

- Deutsche Forschungsgemeinschaft Focus Program SPP 100+ 2388, Grant Nr. 501624329;
- National Science Foundation grants 1623190, HRD-1834620, HRD-2034030, and EAR-2225395;
- AT&T Fellowship in Information Technology;
- program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and
- a grant from the Hungarian National Research, Development and Innovation Office (NRDI).