

Article

Not peer-reviewed version

Hinton Hypothesis and Competition in Artificial Intelligence: A Qualitative Uncertainty Principle of Invisible Hand in a Possible AI-Agent Society

[Yingrui Yang](#) *

Posted Date: 9 June 2025

doi: 10.20944/preprints202506.0653.v1

Keywords: Hinton; duplication; agent; competition; invisible hand; uncertainty; observation; consciousness; reasoning; game



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Hinton Hypothesis and Competition in Artificial Intelligence: A Qualitative Uncertainty Principle of Invisible Hand in a Possible AI-Agent Society

Yingrui Yang

Department of Cognitive Science, Rensselaer Polytechnic Institute, yangyri@rpi.edu

Abstract: The present paper reframes what we call the Hinton Hypothesis, which states that everything of human nature can be duplicated in artificial intelligence. We consider competition to be a significant aspect of human nature. We assume as our working hypothesis that a perfectly competitive AI-agent society is duplicated from human society, like a free financial market. Human agents hesitate between being non-cooperative and cooperative, a hesitation governed by the invisible hand. AI agents observe each other to gain an advantage through more accurate information. By the economic rationality, every agent tends to be the final observer. Thus, the order between observations satisfies the noncommutative law. This is called the qualitative artificial uncertainty principle, which serves as a model of artificial invisible hand.

Keywords: Hinton; duplication; agent; competition; invisible hand; uncertainty; observation; consciousness; reasoning; game

Preface

The present paper reframes what we call the Hinton Hypothesis, which states that everything of human nature can be duplicated in artificial intelligence. We consider competition to be a significant aspect of human nature. We assume as our working hypothesis that a perfectly competitive AI-agent society is duplicated from human society, like a free financial market. Human agents hesitate between being non-cooperative and cooperative, a hesitation governed by the invisible hand. AI agents observe each other to gain an advantage through more accurate information. By the economic rationality, every agent tends to be the final observer. Thus, the order between observations satisfies the noncommutative law. This is called the qualitative artificial uncertainty principle, which serves as a model of artificial invisible hand.

This paper has seven sections: 1. Hinton Hypothesis. 2. Degrees of Knowledge. 3. Hinton Hypothesis and the Turing Test. 4. Competitions in Artificial Agent Society. 5. Invisible hand: Hesitation between the Non-cooperation and the Cooperation. 6. Qualitative Uncertainty Principle as the Invisible Hand. 7. Conclusion.

1. The Hinton Hypothesis

In a recent interview with Guyon Espiner, Geoffrey Hinton claims, "There isn't this magical barrier between machines and people, when we people have something very special machine could never have. We have a very long history as a species of thinking we're special. We thought we are at the center of universe, we thought we made in the images of God. We have all these pretensions. We are not special, and there is nothing about us that a machine couldn't duplicate." Let us reframe this claim as a working hypothesis below.

Hypothesis 1 (Hinton). *We assume as our working hypothesis that everything about the human species could be duplicated by artificial intelligence (machine).*

This hypothesis follows inductive logic. Based on some facts that have been observed, we jump into a hypothesis. In mathematics, Riemann hypothesis has not been solved since 1859. However, taking Riemann hypothesis as granted, more than a thousand mathematical theorems have been established. Not only do these results enrich our knowledge, but also, they advance our mathematical cognition. Thus, we can reasonably expect that Hinton hypothesis will do the same or even more in the domain of artificial intelligence. Overall, scientific hypotheses are necessary as scaffolds for the advancement of science.

2. Degrees of Knowledge

In epistemology, there is a platonic tradition, which says: If one claims knowing a piece of knowledge P , the following three conditions must be satisfied simultaneously. First, P is true. Second, one believes in that P . Third, one way or another, there are some justifications of that P . We may well think of the Hinton hypothesis as a piece of knowledge of Hinton. However, Hinton Hypothesis is only a pre-knowledge to the present author because I am not sure about it. Knowledge can be classified in three levels:

The first level is called pre-knowledge. A working hypothesis or a working definition may serve as our pre-knowledge. We may have certain observations as justifications, we may believe in it, but it has not been demonstrated or proved to be true.

The second level is called meso-knowledge. Taking a scientific hypothesis as granted, those pieces of knowledge are developed by assuming a hypothesis is classified as meso-knowledge.

The third level is called full knowledge, which is commonly agreed and upon knowledge shared by scientific communities.

This paper is based on the Hinton hypothesis, and it aims to contribute a particular piece of meso-knowledge about a possible perfectly competitive AI-agent society.

3. Hinton Hypothesis and the Turing Test

Hinton hypothesis and the Turing test belong to two different paradigms. The Turing test is a behavioral test, while Hinton hypothesis is a larger and richer cognitive test. Hinton not only claims that the machine could reason but that it also could duplicate human consciousness and emotion. Hinton identifies himself as “a psychologist hidden in artificial intelligence”.

The Turing test and the Hinton hypothesis also had different motivations. Turing was influenced by Gödel and Tarski. Gödel published his independent result (known as Gödel incompleteness theorem) of the first-order theory in 1931. In 1933, Tarski proved that the truth cannot be defined as a predicate in the first-order theory. This is the well-known undefinability theorem. Like Gödel and Tarski, Turing was motivated to understand the fundamental limitations of computation. In 1936, he introduced the concept of the universal Turing machine, and asked a foundational question: What can't the machine do? In 1950, Turing revisited his inquiry and posed a new question: Can machines think? In that context, he proposed the now-famous Turing test as a practical substitute for the original question, which he argued was too vague and ill-defined to address directly (Turing, 1950). Today, from many perspectives, the Hinton hypothesis has moved beyond the scope of the Turing test, raising new challenges regarding what it means for machines to exhibit intelligent behavior. Turing looks for the limitations of artificial intelligence, while Hinton looks for the potential of artificial intelligence. They have different faiths.

Hinton can be seen logically as a follower of the functionalism of William James (1989/2017). Particularly, James spent much of his book discussing brain and neural function. From the psychological perspective, neural networks are seemingly rooted in functionalism from the computational perspective. Hinton identifies computational procedures in the LLM neural network as mental acts such as reasoning, consciousness, and emotion. From the philosophical perspective,

Hinton's view is akin to Whitehead's procedural philosophy (1979), which is committed to procedural reality. From the logic perspective, Turing test can be characterized as an existentially quantified statement, while the Hinton hypothesis is described by a universally quantified statement. The latter is a much more ambitious advancement than the former.

4. Competitions in Artificial Agent Society

Competition is a significant character of human society. Competitions happen among individuals, companies, nations, etc. By the Hinton Hypothesis, all of these competitions could be duplicated into the AI-agent society. Hinton deeply worries about how the development of AI might cause humanity to lose control. Hinton claims that AI companies have only short visions and mostly focus on short term benefits, which actually acknowledges that the competitions among AI companies are real. On one hand, Hinton counts on the government regulations to solve the problem; on the other hand, he realized that AI has been nationalized with even more serious competitions.

By the Hinton's Hypothesis, we may assume a perfectly competitive society of AI-agents which is duplicated from human society, such as a financial market. Assume this AI-agent society is a free society that is perfectly competitive. Each agent tries to take more resource from others. In order to do so, each agent needs to observe what other agents are doing; they are watching each other. In other words, each agent must observe what other agents have observed in order to obtain more information that they could take advantage over other agents and benefit from it.

We may assume that this is pushed by economic rationality, each and every agent keeps this observation activity in order to maximize the benefit. We call this game the perfectly competitive game.

5. Invisible hand: Hesitation between the Non-cooperation and the Cooperation

This section is largely duplicated from Yang (2023, 2024). Let us first review the representation of the Nash equilibrium in non-cooperative game theory (Osborne and Rubinstein, 1994). The basic syntactic structure of non-cooperative games is quite simple. Consider n players, where each player i has a set of possible actions $A_i = \{a_{i1}, \dots, a_{im}\}$. Each player establishes a total preference relation, denoted as $\succ_i$. It is important to note that in individual decision theory, a decision maker's preference relation is based on their own set of possible actions. In contrast, in game theory the preference relation of any player can only be established based on what is referred to as the set of action profiles. Considering the possible action sets of all players $A_i (i = 1, \dots, n)$, the Cartesian product can be expressed as:

$$\times A_i = \{(a_1, \dots, a_i, \dots, a_n) | \forall a_i \in A_i\}$$

In this context, each n -tuple $(a_1, \dots, a_i, \dots, a_n)$ is referred to as a situation. In other words, a specified game constitutes a set of situations, and each player must establish their own total preference relation over this set of situations. That is to say, for all players i , each must establish their own $\succ_i$ on $\times A_i (i = 1, \dots, n)$. Once the syntactic structure of non-cooperative games is understood, it is not difficult to grasp its key meta-property, namely the well-known Nash equilibrium. It is important to note that the language of Nash equilibrium requires a separate characterization for each player. Therefore, to reformulate the expression for the n -tuple, we have:

$$(a_1, \dots, a_i, \dots, a_n) = (a_1, \dots, a_{i-1}, a_i, a_{i+1}, \dots, a_n) = (a_i, a_{-i})$$

Here, $a_{-i} = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$. The Nash equilibrium is a specific scenario $(a_i, a_{-i})^*$, namely that of (a_i^*, a_{-i}^*) such that for each player i , and for any $a_j \in A_i, j \neq i$, it holds that:

$$(a_i^*, a_{-i}^*) \succ_i (a_j, a_{-i}^*)$$

The concept of the Nash equilibrium requires some thoughtful interpretation in mathematical terms. In simple terms, it suggests that in a non-cooperative game, each player loses, but all lose equally. It is important to note that the language used to characterize the definition of the Nash equilibrium captures the actions a_i of any individual and the set of actions of all other individuals a_{-i} in the same situation. It is representative of the separation approach, a typical technique in mathematics for characterizing fixed-point problems.

The foundational theoretical framework of the theory of competition in cognitive science remains the Nash framework. Within this framework, a strict mathematical distinction is made between non-cooperative and cooperative games, and the overarching meta-properties of both, namely the Nash equilibrium and the Nash solution, respectively. However, a significant body of behavioral game theory research (Camerer, 2003) highlights a phenomenon where players oscillate between non-cooperative and cooperative games, which can be termed as *fluctuations*. For instance, the classic prisoner’s dilemma, presented in nearly all game theory textbooks, is originally designed as a non-cooperative game. However, altering the game’s conditions—such as increasing the duration of rewards and penalties or allowing repeated play—can lead players to shift from a non-cooperative state to a cooperative one. Note that this phenomenon is exactly as Adam Smith (1776) characterized as the free market, which supposes to be governed by the so-called *invisible hand*. The main purpose of this paper is to find this invisible hand governing the perfectly competitive AI-agent society.

These behavioral fluctuations are directly observable and as such are classic examples of fluctuations. The fluctuations identified by behavioral game theory in empirical studies cannot be adequately explained within the standard Nash framework of game theory (Osborne and Rubinstein, 1994). The underlying causes and corresponding theoretical explanations must be sought in the realm of individual decision-making theory.

To construct a unified theory that decomposes a game problem into decision-making problems for each player, it is essential to translate the formalism of game theory into the formalism of decision-making theory. This requires some technical adjustments. When game theory is cast to address a specific player i , a situation $(a_1, \dots, a_i, \dots, a_n)$ can be rewritten as (a_i, a_{-i}) . We will now make a further revision, transforming (a_i, a_{-i}) into $\alpha_i(\alpha_{-i})$. The rewritten $\alpha_i(\alpha_{-i})$ resembles a function, which is not conventionally within the scope of game theory; however, this is a critical step in bridging the gap between the formalism of game theory and the formalism of decision theory. We will see why this is the case shortly.

The book by Leonard Savage (1972) is recognized as the seminal work in contemporary axiomatic decision-making theory. Below, we will use Savage’s formalism (1972) to characterize the structure of decision-making problems. A decision-making problem is represented as a triplet $\langle F, S, H \rangle$, where F is a set of action functions, S is a set of states, and H is a set of outcomes. For a given action function $f \in F$ and an environmental state $s \in S$, we have $f(s) = h$, $h \in H$. It is important to note that for a specific state s , the value of $f(s)$ is unique. Therefore, in any non-ambiguous context, h can be omitted. For any two action functions f_1, f_2 , we define a preference relation $f_1 \succ f_2$, indicating the preference of f_1 over f_2 . Now, note by comparison that $\alpha_i(\alpha_{-i})$ from the previous paragraph and $f_i(s)$ here are structurally similar. We can treat α_i in the former as the action function f_i in the latter, α_{-i} as the state variable s , and thus transform $\alpha_i(\alpha_{-i})$ into $f_i(s)$.

Fluctuations of AI society originate, in the strictest sense, from the reasoning processes of AI-agents. These reasoning processes are purely within the mind, difficult to observe directly, and are subject to various individual differences. These details fall within the domain of mental decision logic, and what follows is a brief explanation.

Let us first examine language conversion and predicate relationships. Previously, we translated (α_i, α_{-i}) in the formalism of game theory into $f_i(s)$ in the formalism of decision-making theory. Next, we will convert the formalism of decision-making theory into that of reasoning theory (Mendelson, 2015). This involves treating action functions as predicates and state variables as logical variables. That is to say, we transform $f(s)$ into $A(x)$. At this stage, it is no longer necessary to

reference the indices i that originate from game theory and traverse across the individual players. Reasoning is a purely mental process, and the mind is embodied in individuals. Predicates can represent certain unitary properties or n -ary relations.

The first advantage of this predicate technique is that it allows the editing of an option set for a classic decision problem or an action function set for a Savage decision problem. A decision maker may be disinterested in a particular option or unwilling to pursue a certain action function, leading them to abandon that option or action. In other words, the decision maker can establish predicate relationships between options of interest or actions they are willing to take. This represents the most direct logical step in editing a decision problem, carrying significant psychological and cognitive implications.

6. Qualitative Uncertainty Principle as the Invisible Hand

From the above game-theoretic descriptions, the perfectly competitive AI-agent society can be characterized as an n -player game. Furthermore, this game can be seen as a two-player game, denoted as (α_i, α_{-i}) . We can now rewrite α_i as α and α_{-i} as β . Hence, we now have two observers in the game. Let α_β be the observation of α on the observation of β . Let β_α be the observation of β on the observation α . Driven by economic rationality, every AI-agent wants to be the last observer. So, it is not difficult to understand that the order of the appearance of α_β and β_α is sensitive to the result. We have

Proposition 1. *The AI-agent society is competitive. The formula for its competitiveness is:*

$$\alpha_\beta \cdot \beta_\alpha - \beta_\alpha \cdot \alpha_\beta \neq 0$$

Which usually written as $[\alpha_\beta, \beta_\alpha] \neq 0$. This is called a non-commutative relation. In mathematics, a typical non-commutative relation is the matrix multiplication. Heisenberg first discovered that the momentum variation Δp and position variation Δx of the wave function do not satisfy the commutative relation, his mentor Bohr immediately thought of matrix multiplication. Therefore, the Heisenberg picture of quantum mechanics was first called matrix mechanics.

It should be pointed out immediately that market dynamics is currently a qualitative theory, in the semi-dimensional stage. This reflects its social science characteristics, but does not reduce its significance in terms of conceptualization and structure. The form of Heisenberg's uncertainty principle is, $[\Delta p, \Delta x] = i\hbar$, where $\hbar$ is a unit of energy, known as Planck's constant. This shows that energy is not continuous but discrete. In other words, energy is calculated in Planck units. In the atomic energy level model, electrons in low-energy polar orbits need enough units of energy to jump to higher-energy orbits. This is a matter of course for market dynamics. Prices have always been discontinuous; a commodity that costs 10 dollars per piece cannot be bought with 9 dollars. Reviewing the definition of interval in the special theory of relativity, the first term on the right is the energy term; in market dynamics, its meaning is the square of the money speed multiplied by the absolute price. As for bargaining, that is another story.

When writing academic articles, we always need to give keywords, that is, the key concepts to understand the article. Here, the non-commutative relation is a key concept. The key point is that quantization is to find a non-commutative relationship; once a non-commutative relationship is established, the system is quantized (Wang, 2008). This is from the perspective of market language syntax. In addition, from the above analysis, it is not difficult to see that from the perspective of the semantics of market language, the meanings of α_β and β_α are the information obtained by both, and the amount of information is slightly changed to $\Delta \alpha_\beta$ and

$\Delta \beta_\alpha$. From the market competitiveness, we should know that these are two uncertain quantities and they should obey the Heisenberg uncertainty principle. If the accuracy of one of the uncertainties is higher, the accuracy of the other uncertainties will be lower. This is called the market version of

the uncertainty principle. So far, we realize that in market dynamics, markets are quantized phenomena. We have

Proposition 2. *The invisible hand in the market, also known as the quantum version of the invisible hand, is an interaction of competitive observations that satisfies the non-commutative relationship and adheres to the market version of the uncertainty principle.*

It is worth emphasizing that the quantum version of the invisible hand defined above meets a key condition: it is invisible to all market participants. Note that in the previous definition, for any given α_β and β_α , the non-commutative relationship is satisfied. At the same time, for any information $\Delta \alpha_\beta$ and $\Delta \beta_\alpha$, the uncertainty principle is satisfied. This implies that for all AI-agents, they have the same status relative to the quantum version of the invisible hand. This is a kind of symmetry. It is precisely because of this symmetry that the AI society can become a fair society; and only an AI society that is fair to every AI-agent can be conserved and become a sustainable society. Imagine if an AI-agent has a channel to obtain inside information from some governing authority and thus has the opportunity to manipulate the development of AI industry, then this agent can use the information advantage to continuously obtain the necessary resource from other agent's share and destroy the symmetry. In this way, after a long time, others can only withdraw from the competition, making it difficult for the AI society to develop sustainably. The quantum version of the invisible hand is equivalent to what is called the non-observables by T. D. Lee (1988). This reflects a famous theorem, known as the Nöether's theorem: non-observables imply symmetry, and symmetry implies conservation. The reason is that symmetry represents a certain invariant, and in this sense, the AI society can become a conservation system. We can see that Nöether's theorem is not only profound for mathematical physics, but also for dynamics of the AI-agent society, economics and even social sciences in large.

Nöether's theorem tells us that if the AI society is to develop sustainably, symmetry, that is, fairness, must be established among all AI-agents. To achieve this, we must find something that is superior to any market participant. This is the highest principle of the AI society mechanism design and the meaning of the invisible hand of the AI-agent society. Note that this article assumes that the AI society is free, i.e., perfectly competitive. We assume that the perfectly competitive AI society is a closed, independent system, or a conservative field. As we can see, this conservative field satisfies U (1) symmetry. External forces can, of course, destroy the closedness of this conservative field, leading to spontaneous symmetry breaking. In artificial intelligence, we collectively refer to the so-called external forces, like AI policies such as government regulations, capital investments, or energy crises, as AI externalities.

7. Concluding Remarks

Remark 1. *At the present, it is unclear if we should take the Hinton hypothesis as a scientific hypothesis. We have not found a way to disprove it, and at this point, it's hard to imagine the boundary of artificial intelligence.*

Remark 2. *This paper demonstrates that the Hinton hypothesis is useful, which may lead us to address many theoretical issues concerning artificial intelligence.*

Remark 3. *Artificial intelligence is not only a technology but an integration science involving humanities and social sciences.*

Remark 4. *Current AI theories are mostly phenomenological or statistical. AI urgently demands more basic theories beyond philosophical inquiries. Mathematics and theoretical physics have many modeling tools available for use to apply. According to the view of Nobel Prize committee, AI have been counted as a part of physics (Nobel Prize in Physics, 2024). This paper constructed a qualitative version of the uncertainty principle as the model of invisible hand. This method is borrowed from quantum mechanics.*

Remark 5. Artificial intelligence is a rather special domain. It is not only a technology and a science, but also a new prolificacy and thus a business. Hayek has a book (2009) on Denationalization of Money. This book acknowledges that money is nationalized, and that this phenomenon will last for a long time. Hinton seems to dislike AI companies that focus on making short-term benefits. This tells us that the AI industry is largely business-oriented, and that this phenomenon will continue for a long time.

Remark 6. Among many human features, competition is probably the closest to the animal world. In other words, our food chain will inevitably extend to include AI as a new species.

Remark 7. Could an AI-agent society be evolving to a stage against the human society? The observations show that most people are not currently worried about it, except Hinton and others like him.

Remark 8. Would human society take advantage of artificial intelligence for human competitions? The observation shows that most people are currently concerned, except for Hinton and others like him.

References

1. Camerer, C. F. (2003). Behavioral Game theory: Experiments in Strategic Interaction. Princeton University Press.
2. Hayek, F. A. (1990). Denationalization of Money: The Argument Refined. Institute of Economic Affairs.
3. James, W. (1900/2017). Principles of Psychology. Great-Space Independent Publishing Platform.
4. Lee, T. D. (1988). Symmetries, Asymmetries, and the World of Particles. University of Washington Press.
5. Mendelson, E. (2015). Introduction to Mathematical Logic. CRC Press.
6. Osborne, M. and Rubinstein, A. (1994). A Course in Game Theory. MIT Press.
7. Savage, L. (1972). The Foundations of Statistics. Dover Publications.
8. Smith, A. (1776/2023). The Wealth of Nations. Fingerprint! Publishing.
9. Turing, (1950). Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59, 433–460.
10. Wang, Z. X. (2008). Elementary Quantum Field Theory. Peking University Press.
11. Whitehead, A. N. (1979). Process and Reality. Free Press.
12. Yang, Y. (2023). Contents, Methods, and Significance of Higher Order Cognition Study. Advances of Academics (in Chinese).
13. Yang, Y. (2024). Contents, Methods, and Significance of Higher Order Cognition Study. (in English). Preprint,

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.