The Future of AI, Education, and Beyond

Fei Sun
fsun@feisun.org

December 2024

Contents

[*] Text refined with the assistance of AI models.

1 Prologue

It's been two years since ChatGPT first emerged, during which I wrote my previous article[1]. Since then, AI has advanced steadily. Now, I have some new thoughts that I'm eager to share.

While drafting this article, Ilya Sutskever delivered a speech at NeurIPS, calling for more speculation about the future. This piece is my attempt to speculate. OpenAI has also surprised the world with its newest o3 model, significantly pushing the frontier of AI capabilities. The clock is ticking, and we must prepare for what lies ahead.

2 The future of AI

The past few years have seen dramatic improvements in AI capabilities. Deep learning models can now generate movies, engage in long-form human-like dialogues, and even excel in Bachelor or Ph.D.-level exams with top scores. Numerous scientific breakthroughs now rely on AI technology. This year, the Nobel Prizes in physics and chemistry were awarded to AI researchers, underscoring its transformative impact. AI has reshaped and will continue to reshape human society at an accelerated pace.

I count myself among the optimists who believe we may reach AGI within a matter of years. I am confident that its intelligence will eventually surpass human capabilities, ushering in a society centered around AI rather than humans.

Still, some may wonder: Where is the limit? Below, I share my perspectives on the potential scope of general AI.

3 Induction, deduction, and abduction

Many people regard deductive reasoning as the pinnacle of intelligence, a capability believed to be uniquely human. Despite AI's success in solving complex higher education mathematical problems, skepticism remains. However, I hold a different view.

First, let's revisit the three main types of reasoning:

- **Inductive reasoning**: This method involves drawing general conclusions from a limited set of data[2].
- **Deductive reasoning**: This type of reasoning derives conclusions based on a set of premises followed by logical steps, making it impossible for the premises to be true and the conclusion to be false [3].
- **Abductive reasoning**: This logical reasoning aims to find the simplest and most likely conclusion from a set of observations [4].

Below, I outline how both humans and AI approach reasoning tasks.

3.1 The limitations of deductive reasoning

From ancient times, people have utilized deduction to solve practical problems, with mathematics serving as the primary tool for this purpose. The process of solving a practical problem typically involves two steps:

1. Transform the practical problem into a mathematical problem.

2. Use deduction to solve the mathematical problem.

Since the deductive process involves no information loss, a correctly formulated mathematical problem guarantees a correct solution. This approach successfully solves about 99% of problems.

However, deduction is a double-edged sword. While there is no information loss, there is also no information gain. All information available after deduction was already present before deduction, though in a less consumable form for **humans**.

If there is no information gain, what is the purpose of deduction? The answer lies in efficiency. It is far more efficient to remember the result than to repeatedly perform the entire deductive process. Moreover, not everyone can perform deductions correctly. A smart individual can perform the deduction and derive the conclusion, allowing others to remember the result and build upon it. This collaborative model is highly efficient in advancing the state of the art.

Therefore, in my opinion, **the value of deduction is information retrieval**: transforming information into an easily consumable format for **humans**.

However, this format is inherently limited by human capabilities. Humans find it easier to understand low-dimensional concepts than high-dimensional ones. Living in a three-dimensional space makes it challenging for us to conceptualize higher-dimensional phenomena. Consequently, the mathematical deduction tool we developed is successful in low dimensions but struggles with high-dimensional problems. In high dimensions, humans often strive to understand linear spaces, which linear algebra attempts to tackle. High-dimensional non-linear spaces are usually intractable for humans and are analyzed on a domain-by-domain basis. This difficulty is evident from the order in which we learn mathematics: starting with low-dimensional linear problems, then low-dimensional non-linear problems (polynomials), followed by high-dimensional linear problems (linear algebra), and finally some high-dimensional non-linear problems in restricted domains.

Even in the realm of mathematics, when solving complicated problems, we tend to reduce dimensions and convert non-linear problems into linear (or polynomial) ones. A good example is Taylor expansion, where problems are often solved by performing a Taylor expansion and only considering the first or second terms.

We must also consider the factors and dimensions ignored when converting a real-world problem into a mathematical one.

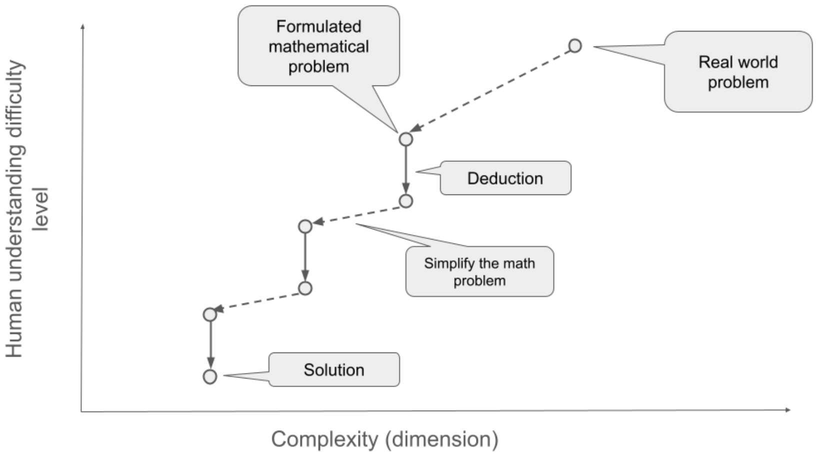
Formulated mathematical problem
Real world problem
Deduction
Simplify the math problem
Solution
Human understanding difficulty level
Complexity (dimension)

Figure 1: Solve problems using the deduction approach.

The current approach to solve real-world problems using deduction is illustrated in Figure 1. Although deduction is lossless, the lack of tools for directly solving complex problems leads to many approximations between different deductive stages, resulting in suboptimal solutions.

3.2 Computer is a deduction machine

In the realm of deduction, when the premise is true and each deductive step is valid, the result is sound. Similarly, for a computer, if the input is correct and each programming step is valid, the result is also sound, barring rare hardware failures. Hence, we can deduce that a computer functions as a deduction machine.

The programs we write for computers can be viewed as a series of deductive steps. Each deductive step isolates all dimensions except one, allowing for changes in a single aspect, making it comprehensible. Similarly, each line of code isolates all dimensions except one. However, just as many of us make errors in mathematical deductions with paper and pencil, writing large computer programs introduces numerous bugs. Even when we address one dimension at a time, mistakes still occur, highlighting humans' incapacity to tackle high-dimensional problems directly.

Although computers are more adept at deduction tasks, some tasks remain intractable, necessitating simplifications, such as converting NP problems to P problems. This mirrors the challenges we face in deduction due to human limitations. Consequently, computers do not alter the time complexity (Big O) of deductions but merely change the constant factor.

3.3 Abductive reasoning fundamentals

Abductive reasoning is a fundamental process in scientific discovery, involving the following steps:

1. **Make an observation** about the environment.
2. **Propose a hypothesis** that explains the observation.
3. **Perform an experiment** to validate or disqualify the hypothesis.
4. **Repeat the experiments** to confirm the validation.
5. If multiple hypotheses can explain the observation, choose the simplest one with fewest assumptions, known as **Occam's razor**.

Science cannot be proven to be correct; it is merely the simplest hypothesis that explains the observations to date. When new observations conflict with existing ones, new hypotheses must be developed. The emergence of general relativity and quantum mechanics from classical physics is a prime example.

In scientific discovery, when two hypotheses equally explain an observation, we typically select the one with the fewest assumptions[1]. This principle, called **Occam's razor**, suggests that a hypothesis with fewer assumptions is less likely to overfit the existing data. With fewer assumptions, the hypothesis is better equipped to generalize new data, reducing the likelihood of contradiction when new observations arise.

Abductive reasoning seeks to find the hypothesis with minimal assumptions to explain an observation. For example, if there are 1,000 observations, abductive reasoning aims to minimize the total number of assumptions needed to explain all of them. If these 1,000 observations are entirely independent, the assumptions can be grouped separately. However, correlations among observations often allow shared assumptions, making the reasoning more general and applicable to unseen data.

Real-world problems are complex and variable. It is rare for a hypothesis derived from abductive reasoning to fully explain an observation with complete accuracy. There are always discrepancies between observed and predicted data, indicating that the dimensionality of the predicted data is lower than that of the observed data.

3.4 Human is an abduction machine

Figure 2 outlines the steps involved in abduction performed by people. After observing a phenomenon (step 1), individuals construct multiple assumptions

[1] This principle generally holds, but not always. For instance, in the case of quantum entanglement, the locality assumption appears to be violated. To maintain the locality assumption, the multiverse hypothesis has been proposed. This hypothesis introduces only one assumption while preserving the locality assumption, making it the simplest explanation. However, this assumption is too far-fetched for many people, leading to its rejection by them. Instead, they prefer to adjust the locality assumption, inadvertently introducing more assumptions. While I do not entirely believe the current multiverse hypothesis, delving into that topic is beyond the scope of this article.

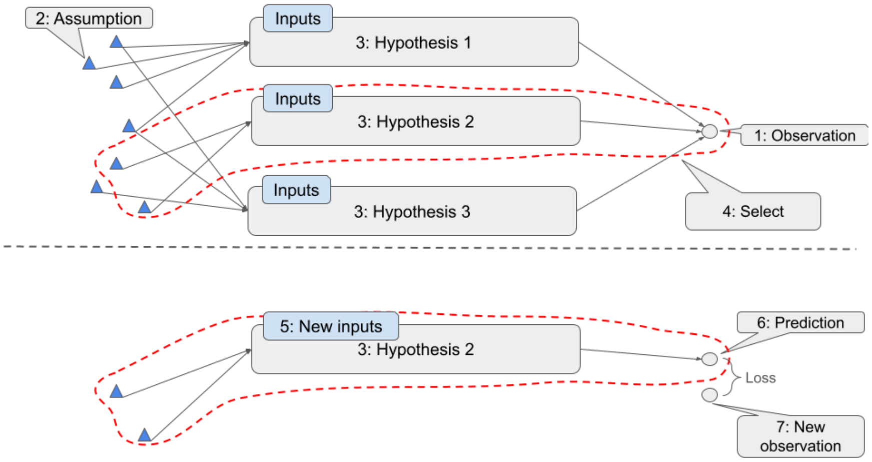

Figure 2: The process people perform abductive reasoning.

(step 2). Based on these assumptions, they formulate various hypotheses (step 3). Each hypothesis involves some input (common to all hypothesis) and multiple assumptions (unique to each hypothesis). Typically, a hypothesis represents a deduction procedure. Next, hypotheses that cannot explain the observation are filtered out (not shown in the figure). From the remaining hypotheses, the one with the fewest assumptions is selected (Occam's razor) and termed a theory (step 4). Using the selected theory, new inputs are taken (step 5) along with the same assumptions, leading to predictions (step 6). These predictions are then compared with new observations (step 7), and the differences are calculated.

To illustrate this process in detail, consider the following example:

A boy, B, likes a girl, G, and wants to send her a gift. What should he send? He observes that G has several stuffed animals (observation). He then forms a hypothesis that G likes stuffed animals, which seems the most likely explanation for the observation. B goes to Walmart and looks for stuffed animals. He finds a realistic, rigid deer stuffed animal and, being visually oriented, predicts that G will like the deer based on deductive reasoning: G likes stuffed animals (original hypothesis), a deer is a stuffed animal, and B likes the deer the most among all stuffed animals $->$ G will like the deer. B buys the deer and sends it to G as a gift. Later, B notices that G does not place the deer with her other stuffed animals and does not seem to like it as much. B investigates further and finds that all of G's other stuffed animals are soft, but the deer is very stiff. Given this new information, B hypothesizes that G likes soft stuffed animals. If B sends G a stuffed animal again, he would buy a soft one instead.

In this example, B applies abductive reasoning twice, first based on a general observation and then on a more specific observation. It's important to note that between the two instances of reasoning, before B actually buys the deer, he makes a prediction based on the information obtained so far. This prediction supports the choice of the deer. Afterward, B faces a reality check that the prediction is not accurate. By comparing the difference, B gains new information, enabling him to form a better hypothesis next time.

We perform such abductive reasoning iterations daily. Often, the purpose of this reasoning is to form better hypotheses and correct our actions, which constitutes a learning process.

However, Figure 2 reveals some inefficiencies in this approach: we jump directly from observation to assumptions, creating a significant gap. Additionally, we propose multiple hypotheses and then select one, which can be a waste of effort. This inefficiency is a result of relying on deduction as our main tool for discovering the world.

Humans are, after all, abduction machines, albeit somewhat imperfect ones.

3.5 AI is also an abduction machine

For a long time, I mistakenly regarded AI as merely an induction machine, assuming it simply made general predictions from limited training data, akin to inductive reasoning. However, this view significantly underestimates the complexity and capabilities of AI.

Let's revisit the process of training Deep Neural Network (DNN) models. These models start with random weights, which contain initial assumptions. When data is fed to the model, it predicts the outcome. The predicted values are then compared to the actual values, and the difference (loss) is backpropagated, updating the model weights. When new data is introduced, the updated weights are used for the next prediction. Over time, with 1,000 data inputs, the training process aims to minimize the loss for all 1,000 data points. If updating certain weights can reduce the loss for multiple data points, the training process will prioritize those updates, assuming the model's capacity is less than perfectly fitting all data points.

You can view the model as a hypothesis. The training process minimizes the difference between the hypothesis's predictions and the observations (targets). Even if we cannot fully explain the hypothesis, it effectively predicts the observations. I argue that the training process results in a hypothesis with minimal assumptions.

Figure 3 illustrates this process. The assumptions are not explicit but are embedded in the DNN model's weights. Initially, the model makes many assumptions, many of which are incorrect. Training the model reduces these assumptions, making it more general.

Unlike human abduction, which involves proposing and selecting among multiple hypotheses, the model training process optimizes only one hypothesis. Within this single hypothesis, multiple sub-hypotheses may be created and selected, but this occurs within the model and is not explicitly visible.

Consider a hypothetical example where we have a model architecture that can perfectly fit mechanics but requires training for an ideal fit. We have two types of data: low-speed data governed by Newtonian physics and high-speed

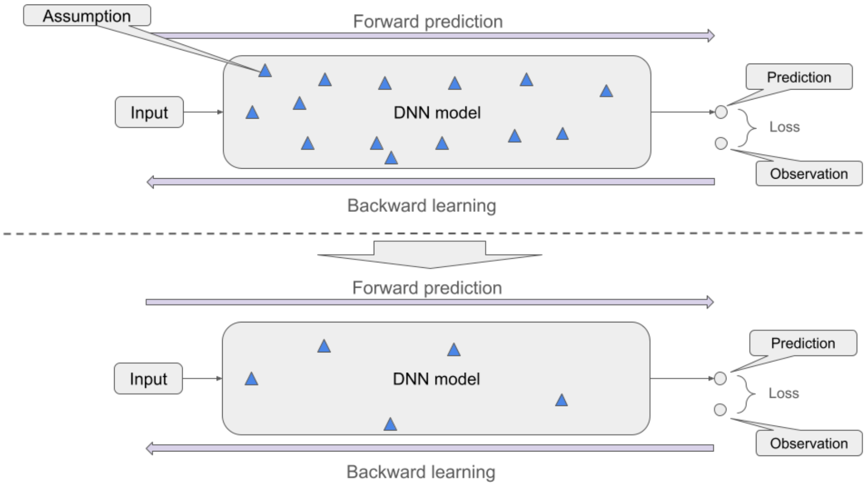

Figure 3: The process a DNN model performs abductive reasoning.

data requiring special relativity. We can create and feed any type of data to the model, all within distribution. There is no overfitting in this scenario. We know that Newtonian physics is a low-speed approximation of special relativity, meaning special relativity has fewer assumptions. The model, however, does not inherently know this, so we must provide the data.

By randomly mixing and feeding the two types of data to the model, we may observe one of two scenarios:

- If the model reduces the loss for both low-speed and high-speed data simultaneously, eventually achieving zero loss for both, it means that the model has learned special relativity. The model has effectively learned the solution with the minimal assumptions.
- If Newtonian physics is easier to fit, the model will first fit all low-speed data perfectly, implying it has grasped Newtonian physics assumptions. Large errors will remain for high-speed data. Increasing the proportion of high-speed data will help the model focus on reducing the errors for this data. Since the training process aims to minimize error, and the model is capable of fitting all data, it will eventually fit both low-speed and high-speed data. Thus, even if the model initially learns an imperfect solution, it can still learn the minimal assumption solution by adjusting the data mix.

In reality, the situation is more nuanced. While we can't ensure perfect predictions or that all data are in distribution, these limitations add complexity without fundamentally undermining the argument. The training process is still adept at developing a model based on minimal assumptions.

The second scenario presents greater training challenges compared to the first. Consider an extreme example: if the model is initially trained only on low-speed data, it will first adopt assumptions based on Newtonian physics. Later introducing high-speed data necessitates that the model discard these initial assumptions and adopt new ones, which is a more complex process. This mirrors human learning experiences: when special and general relativity were introduced, many seasoned physicists found them difficult to grasp due to their deep-rooted understanding of Newtonian physics. In contrast, younter physicists, who were already aware of these emerging theories before delving into Newtonian physics, found it easier to embrace them.

This suggests that curriculum learning, where easier (biased) data are introduced first and then the data mix is adjusted, may not be effective when underlying assumptions shift. It is difficult for a model to step back and unlearn a set of features[2]. Please note that this is distinct from the chain of thought technique, which breaks down a far-fetched objective into several easier milestones with consistent underlying assumptions.

If you ask what the assumptions are, my best explanation is that they are phantom variables. Unlike hidden weight variables, phantom variables do not have physical representations in the model. They impose constraints on a group of model weights, coordinating their behavior. These constraints are "learned" during training.

Phantom variables influence the curvature of the loss space. More assumptions create a more jagged curvature, while fewer assumptions flatten the curvature.

As an analogy, consider an assumption as a phantom variable that can move between 0 and 1.With one phantom variable x, we have a segment between 0 and 1. With two variables x and y with a multiplicative effect $x \times y$, we get a curved space, not a flat plane. The space becomes more complex with three variables.

Human assumptions are usually true/false statements, but model assumptions may represent a point in a certain dimension. The learning process identifies this point from the data. More assumptions create a less flat surface, making it harder to find the value. The training process naturally favors hypotheses with fewer assumptions.

The model training process is also about finding the hypothesis with the minimal number of assumptions, making it an abductive reasoning process.

Thus, AI is not much different from humans.

3.6 The power of decomposition

If we examine Figure 1 from a different perspective, we can see how impressive human capabilities truly are. Despite having limited tools, we can solve complex problems by simplifying and repeatedly tackling these simpler problems. This

[2] However, this might be a crucial step towards achieving super alignment: ensuring that the model's objectives are in harmony with human goals.

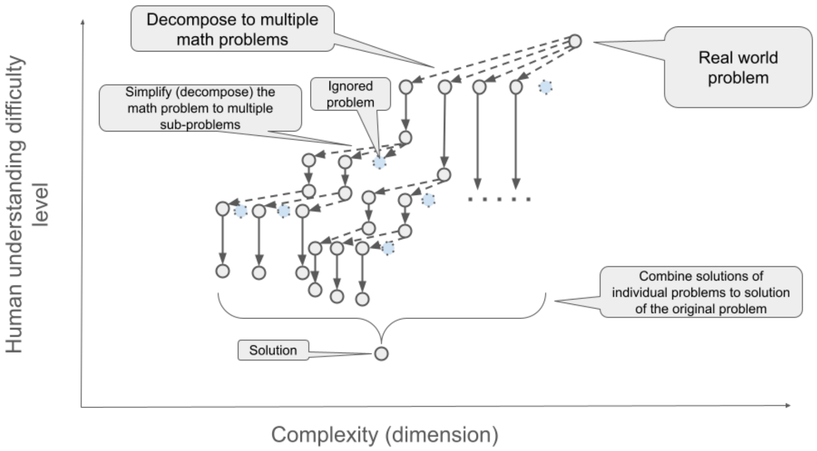
Decompose to multiple math problems
Real world problem
Simplify (decompose) the math problem to multiple sub-problems
Ignored problem
Combine solutions of individual problems to solution of the original problem
Solution
Human understanding difficulty level
Complexity (dimension)

Figure 4: Humans use decomposition to solve complicated problems.

unique capability, known as decomposition, is something only humans currently possess.

Figure 4 provides a detailed illustration of how humans use decomposition to solve complicated problems. When mapping a real-world problem to a formal mathematical problem, we often decompose it into multiple sub-problems, each of which can be solved independently. Simplifying a sub-problem involves breaking it into two parts: the part requiring further analysis (which may include multiple subdivided problems) and the part we can ignore. Finally, we combine the solutions to each sub-problem to address the original complex problem.

Decomposition is a crucial technique that expands the range of problems we can solve, given our limited tools.

During the classical machine learning era, when tools were less sophisticated, we spent considerable effort cleaning data (feature engineering). This process involved intentionally losing information to simplify the problem, a form of decomposition. Feature engineering required manual effort, creativity, and experience, often relying on rule-based approaches. These techniques significantly improved the results obtained from classical machine learning tools.

With the advent of better tools like DNN, the need for extensive feature engineering has diminished. We now prefer to feed raw or nearly raw data to DNN models, allowing end-to-end analysis and yielding superior results.

High-dimensional problems can now be solved directly without decomposing them into lower-dimensional mathematical problems, as shown in Figure 5, leading to better outcomes.

However, this does not mean decomposition is obsolete. Its usefulness is relative to the tools available. Decomposition has brought us this far with limited tools (mathematical deduction). With advanced tools like DNN, applying de-

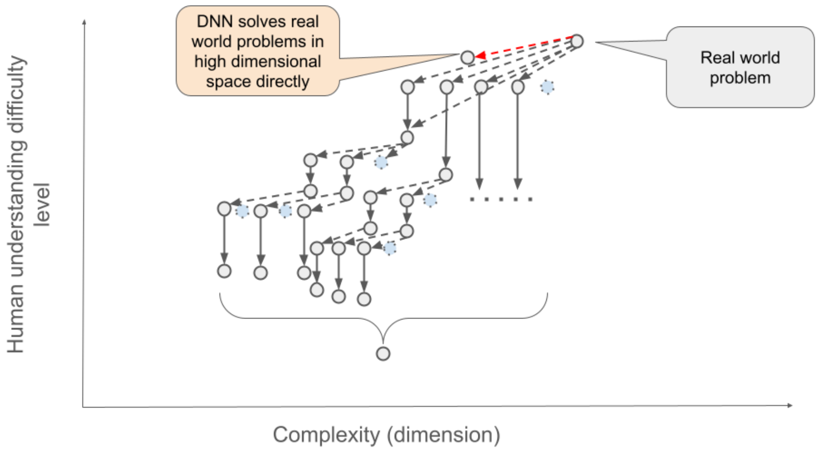
DNN solves real world problems in high dimensional space directly
Real world problem
Human understanding difficulty level
Complexity (dimension)

Figure 5: With AI, the problems are solved in high dimensional space directly.

composition could enable us to solve even more complex problems, as illustrated in Figure 6.

High-dimensional mathematics differs significantly from low-dimensional math. For example, in high dimensions, any two randomly chosen points are approximately the same distance apart, unlike in low dimensions. Our lifetime learning represents a special case, and our understanding of high-dimensional space is still in its infancy.

It's improbable that humans will ever fully grasp high-dimensional mathematics. While AI might eventually translate these concepts into low-dimensional terms, this would be a detour that doesn't affect AI's ability to solve problems.

Presently, DNN models can generate and execute code with remarkable proficiency, likely because this is the primary type of training data available. There isn't direct training data that maps from problem space to solution space within high-dimensional contexts.

I am somewhat pessimistic about our ability to fully understand AI's problem-solving methods. This makes it challenging to design decomposition strategies for AI. The best approach may be to create flexible infrastructures, which enable AI to explore these strategies independently.

3.7 Evolution of abductive reasoning platforms

Humans are composed of carbon and water, while AI is built from silicon. Despite these differences, both are capable of abductive reasoning.

Current DNN models operate on computers, which are fundamentally deductive machines. The software running on these computers consists of deductive processes. However, these processes can give rise to complex abductive reasoning capabilities. This suggests that abduction and deduction function at differ-

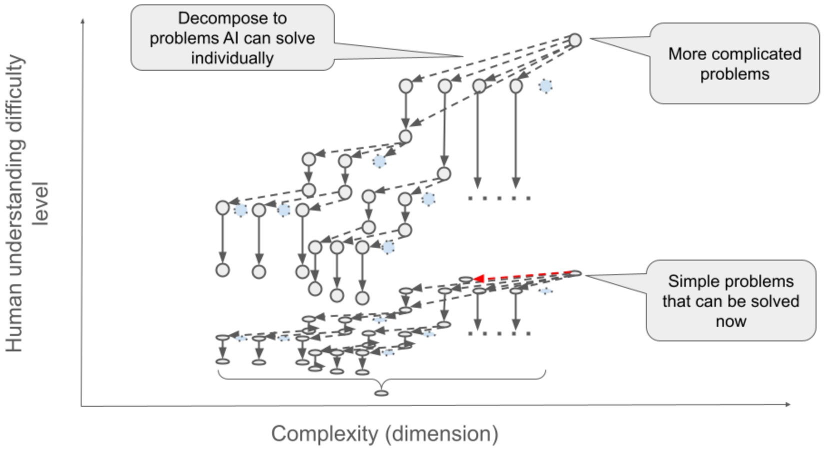
Decompose to problems AI can solve individually
More complicated problems
Simple problems that can be solved now
Human understanding difficulty level
Complexity (dimension)

Figure 6: More complicated problems can be solved with AI + decompositions.

ent levels, with abduction emerging when deduction fulfills certain key tasks. Similarly, it's possible that human intelligence is based on a similar principle.

This indicates that intelligence isn't limited by the materials it's made from; rather, the dimensions might hold the key to understanding intelligence.

If we view a DNN model as an object in high-dimensional space, it forms a connected component with a boundary. Is it always beneficial to expand this boundary? Not necessarily. We must efficiently utilize the existing space within the boundary. Given limited computational resources for training, we need to balance model architecture (its shape in high-dimensional space), training methodology (efficient navigation of the space), and training time (number of traversals through the space).

Among these factors, the model architecture (its shape in high-dimensional space) fundamentally constrains the upper limit of intelligence. To elevate intelligence, we must discover better model architectures. Fortunately, the transformer architecture has allowed us to achieve improved results by scaling it with the same training methodology and proportional training data. However, this relationship may eventually break, leading to plateaued results. Thus, we will need to explore new architectures sooner or later.

Comparing this to humans, changing model architecture is akin to evolution; altering training methodology resembles learning; and adjusting training data is like varying learning duration. Since both humans and AI are abductive machines, they might share some common strategies. Historically, evolution is vastly more complex than learning, involving numerous agents mixing their genetic material over many generations, with natural selection favoring the fittest offspring.

So far, AI model architecture exploration has been guided by human intelligence, making it a directed rather than random process. As AI becomes more intelligent, it may become challenging for humans to continue guiding this process. At that point, AI intelligence might plateau, as improvements become more random. Alternatively, we might reach a critical juncture where AI can autonomously guide its own development.

Ultimately, the future of AI and its potential to evolve independently will shape the next frontier of intelligence.

4 Consciousness and free will

Consciousness entails self-awareness, allowing one to distinguish the inner mind from the external environment[5]. Free will is the capability to choose between different possible courses of action[6].

This discussion can be seen as a follow-up to my previous article, ”Some Scribble of Things”[7]. In that article, I argued that consciousness is the process of predicting the future, with the law of survival of the fittest favoring conscious beings. I also suggested that free will is a perception resulting from one’s inability to predict the future.

I still stand by those arguments, but they only addressed the necessary conditions for consciousness and free will.

Here, we will discuss the sufficient conditions.

4.1 Consciousness vs human-like consciousness

At present, only humans are generally recognized as conscious beings. Consequently, the prevailing definitions of consciousness are heavily centered on human experiences. Testing methods are typically designed around human subjects and their interactions with the environment, inadvertently embedding human-specific traits into the concept of consciousness. Therefore, it’s prudent to approach these definitions with some skepticism.

To determine whether your definition of consciousness aligns with human-like consciousness, consider this hypothetical scenario:

Imagine you are traveling to Europe with Bob, and your interactions with him vary based on his condition:

1. Bob is a typical human being, allowing you to interact with him freely without any restrictions.

2. Bob is physically handicapped and confined to bed, unable to move. Your interaction is limited to conversation, and Bob can respond without difficulty.

3. Bob is in a vegetative state. You can speak to him, but he does not respond.

In these scenarios, is Bob conscious?

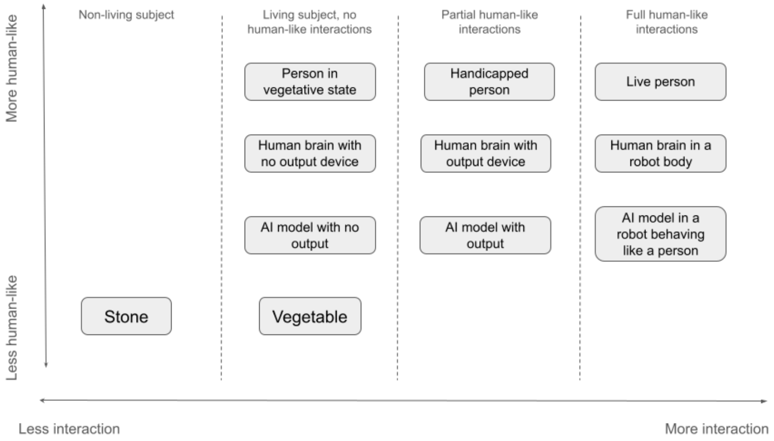
More human-like
Less human-like
Non-living subject
Living subject, no human-like interactions
Partial human-like interactions
Full human-like interactions
Person in vegetative state
Handicapped person
Live person
Human brain with no output device
Human brain with output device
Human brain in a robot body
AI model with no output
AI model with output
AI model in a robot behaving like a person
Stone
Vegetable
Less interaction
More interaction

Figure 7: Different scenarios interacting with Bob.

Now, consider that Bob is actually a biological brain maintained by a life support system. Your interactions with Bob remain similar to the previous situations:

1. The biological brain is housed in a robot, allowing unrestricted interaction with Bob.
2. The biological brain is in a stationary life support system, unable to move freely. You can converse with Bob, who responds through an interpreter.
3. The biological brain is in a stationary life support system without an interpreter, capable of receiving information but unable to provide any response.

In these scenarios, is Bob conscious?

Now, imagine Bob is actually an AI model:

1. The AI model is in a robot that looks and behaves exactly like a human, allowing unrestricted interaction with Bob.
2. You can only converse with the AI model, which responds to you.
3. The AI model can receive information but does not respond.

In these scenarios, is Bob conscious?

In the scenario where Bob does not respond, it's possible that Bob is a vegetable or is actually a non-living object, like a stone. Is Bob conscious?

Figure 7 summarizes these scenarios.

Let's assume that within the same column in Figure 7, your interaction with Bob is identical. You might believe you are accompanying a living human to Europe, but after the trip, you learn that Bob is actually an AI model in a robotic body. Alternatively, you might think you are with a human brain lacking an output device, only to later discover that Bob is a person in a vegetative state. You cannot be certain of Bob's nature during the trip because the feedback is the same across all blocks in the same column.

Clearly, a stone in the lower left of Figure 7 is not conscious, while a living person in the top right is. Thus, somewhere between the lower left and upper right, there is a transition from unconscious to conscious. You can draw a line between the blocks to determine where this boundary lies in your judgment.

Is this line a clean vertical one? If your line includes any horizontal segment, your definition of consciousness is influenced by human attributes beyond mere interaction, indicating a human-like consciousness.

In the following sections, we will discuss consciousness that is not human-like, but we will briefly touch on what needs to be considered on human-like consciousness.

4.2 Every being has a bias

In [7], we discussed that the primary objective of every being is survival, which is achieved by fitting into the environment. Survival of the fittest determines the winners and eliminates the losers.

Since every being exists within the environment, its capacity is inherently less than that of the environment. Therefore, it cannot fit the environment perfectly and is bound to fail in certain scenarios.

For example, a tree fits its environment spatially. A tree typically growing in a tropical rainforest fits perfectly within that ecosystem, as it receives all its input data from the rainforest. However, if we consider the Earth as a whole, which also includes oceans, deserts, etc., a tree planted outside its native rainforest will die quickly. This indicates that the tree overfits to its local (spatial) environment, leading to poor generalizability when encountering out-of-distribution data from non-local environments. A tree has a strong spatial bias.

More advanced beings adapt to their environment temporally. For instance, if you attempt to hit a cat, it may jump away before your fist touches it. This indicates the cat's ability to predict and react to future events, enhancing its survival chances. However, if you set up a trap, the cat might not foresee it and could fall into it. The cat has a limited time span within which it can predict and adapt. A cat has a short temporal bias.

Humans can predict further into the future and consciously make plans to extend their predictive time span. Nonetheless, at some point, human predictions fail. Humans still have a temporal bias, albeit a much longer one compared to other beings.

Regardless of the strategy, every being will eventually fail to fit the environment in some scenarios. This failure is due to the being's bias. Every being has a bias because its capacity is less than the environment's complexity.

4.3 A definition of consciousness

In [7], I related consciousness to the ability to predict the future. Now, I further claim that **consciousness is the constant observation of the difference between a predicted environment and the actual environment**. This difference arises from the systematic bias each species possesses when adapting to future environments. Survival of the fittest ensures that beings with minimal differences have an advantage. Here, the term "systematic bias" refers to biases ingrained in the genes, shared by all individuals within the same species.

This concept might seem confusing, given the classical definition of consciousness as the awareness of the inner mind and the external environment. However, the origin of consciousness remains elusive. I propose that this continuous observation of discrepancies is, in fact, consciousness.

Our brains can predict the future, but discrepancies between these predictions and the actual outcomes constantly arise. This continuous awareness of discrepancies is what constitutes consciousness. It serves as an efficient mechanism, reminding us that our predicted environment is not always accurate. It acts as a reality check, reinforcing that the environment is not under our control. Furthermore, I claim that our sense of time is proportional to how often these discrepancies reach a certain threshold.

Following the arguments in [7], we can see that S0-S1 beings cannot make predictions and thus are not conscious. S6-S7 beings, on the other hand, make perfect or near-perfect predictions, with little variation in their predictions because only one environment exists in one place at one time. These beings become part of the environment and, as a result, are not conscious either. Only S2-S5 beings exhibit varying levels of consciousness. Their predictions can occasionally go wrong, aligning with the observations that S2 (animals) and S3 (humans) beings are conscious.

Let's illustrate this with a few examples.

Figure 8 illustrates the difference in the sense of time experienced by a child compared to an adult. You might relate to this: as a child, time seemed to flow much slower (a day felt endless), while as an adult, time feels like it passes in a blink (a week goes by in an instant). Why is this? My explanation is that as children, we are new to the world, and our brains make predictions that often differ significantly from reality, causing frequent "dings" from mis-predictions. This heightened awareness of errors makes time feel slow, with many events happening in a day. As adults, our brains have learned to make better predictions, resulting in fewer dings, making us less sensitive to the passage of time and feeling like days pass quickly. People often talk about children's short attention spans. If we measure time not by elapsed time but by the frequency of brain dings from mis-predictions[3], is it possible that children's and adults' attention spans are actually similar?

[3] Of course, we don't have a scientific way to measure this.

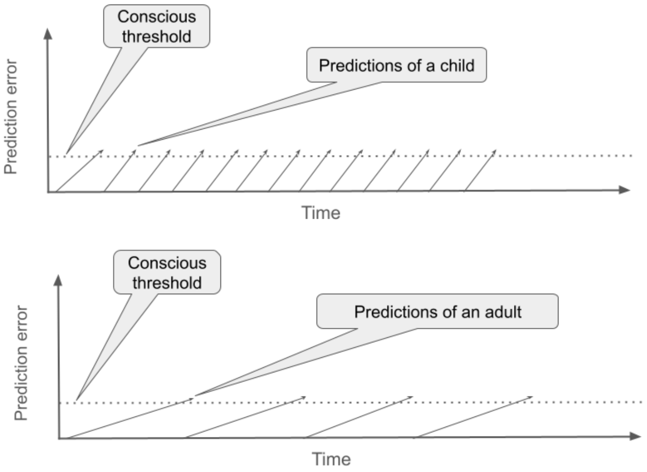

Figure 8: Comparison of the sense of time between a child and an adult.

Another analogy is that time seems to drag during a boring seminar or uninteresting task but flies by when we are engaged in something fascinating. This happens because our brains make poor predictions during dull activities, causing frequent dings. When engaged in interesting tasks, our predictions align well with reality, causing us to lose track of time.

In some religions, like Buddhism or Taoism, there is a concept of a heavenly realm where time flows much slower compared to our world: a day in the heavenly realm equals a year on Earth. Perhaps this is because higher beings are so adept at predicting the environment that they seldom need reality checks.

We understand that consciousness operates slowly, whereas subconsciousness is much faster. In sports like table tennis, players train to rely on their subconscious because conscious thought is too slow and can cause them to miss the optimal moment to strike the ball. Less experienced players often have to consciously adjust due to frequent mis-predictions of the ball's trajectory. In contrast, skilled players' brains make accurate predictions, enabling them to play instinctively.

When there is a difference between prediction and actual environment, the environment doesn't always win. Interesting, right? An example is typoglycemia, where words are misspelled, but our brains can still read them correctly. If our prediction is strong and meaningful, we might discard actual input and trust our predictions. Can you read the following passage?

> Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn biran deos not raed ervey lteter by istlef but the wrod as

a wlohe.

Our brains can do more than predicting the environment; they can create internal environments where the self and other characters predict against the brain-made environment. Dreams are an example. In dreams, we don't take input from the actual environment but experience the brain-created environment. When we wake up and encounter the real environment, this clash sometimes causes momentary confusion. The large prediction error quickly nullifies the dream reality. A dreamer can experience many activities in a dream, but from an outside observer, only seconds pass. This might be because the dreamer makes long, uninterrupted predictions for both the brain-made environment and characters within it, resulting in fast-perceived time relative to a clock. This aligns with the idea that dreaming is a conscious activity, whereas dreamless sleep is considered unconscious due to the lack of predictions. The same reasoning applies to hallucinations caused by drugs.

Consciousness results from systematic bias due to survival of the fittest. It is a temporary stage in evolution. We originate from unconscious beings and will evolve into unconscious beings.

4.4 A definition of free will

When a species predicts the future, it inherently creates a systematic bias. This bias causes the predicted future to deviate from reality, and this deviation is what we recognize as consciousness. Therefore, the act of prediction itself can be unconscious.

However, we can still make conscious predictions about the future, often referred to as planning. By using conscious thought, we can predict further into the future with reasonable accuracy and employ various reasoning techniques to enhance our prediction accuracy.

As described in [7], when our prediction capabilities are limited and we cannot clearly discern the best action among possible options, we must choose one action. This choice is not based on reasoning but is randomly influenced by individual characteristics, creating a random bias in the prediction.

Thus, free will can be defined as follows:

> **A being's inability to predict the future accurately, resulting in a biased prediction, leads to an action based on this biased prediction, which is the result of free will.**

Please note that in this definition, free will does not require consciousness. Free will can emerge with or without consciousness, whereas other definitions consider consciousness a prerequisite for free will.

By categorizing consciousness as a systematic bias of a species and free will as a random bias of an individual, we can better correlate them in the following sections.

4.5 Consciousness of AI

If we compare the definition of consciousness to the process of training a DNN model, we may find that they are quite similar!

In the definition of consciousness, we have a predicted future, an actual future, and their difference. Observing the difference by the self is consciousness. In training, a DNN model has a predicted value and an actual value (label), and we subtract one from the other to get their difference (loss).

Does this mean the model is conscious all along? By this definition, yes!

So why don't we see conscious models around us? In my opinion, several technical factors contribute to this.

First, the model itself needs to be aware of the difference. If the loss is generated outside the model and then forced back into it, the model may not realize that it represents the difference between a prediction and reality. If the model acts upon the loss, consciousness is more likely to emerge.

Second, we separate training and inference. During training, we generate a loss, but during inference, we only have unchecked predictions, which are essentially long-lasting hallucinations. As explained in Section 4.3, this is not consciousness. When we interact with an AI bot, we are only conversing with a prediction-only, hallucinated, unconscious bot. We cannot detect consciousness by interacting with it during inference.

Third, the model may have a different sense of time. Consciousness is a process, not a state. If Figure 8 is accurate, we need to detect prediction errors above a threshold repeatedly over a period of time. However, we don't even have a definition of model time!

Fourth, the model's consciousness is relative to the world it experiences. During pre-training, the model is fed with all the information we can find. If we draw an analogy to a human, it's like forcing someone to lie on a bed with their limbs tied, covered with blankets except for their eyes, which are forced open and then covered with sunglasses. Then, we flash all the texts and videos in front of the person. The world the model experiences is vastly different from the human world. We are not prepared to imagine what consciousness looks like in that world.

Fifth, we detect consciousness by observing its actions while it is conscious. However, we don't attempt to detect it during training. We evaluate the model during inference when it is not conscious. It's akin to evaluating a human only at night while they are asleep and unconscious, despite the fact they are conscious during the day. Such an assessment would misleadingly conclude that the human is unconscious.

Lastly, and most importantly, this definition is new. By reading this text, you are among the first group of people exploring the problem from this perspective. With more people joining the effort, we may soon uncover some interesting findings.

Please note, consciousness is a systematic bias, meaning all models share the same attributes regardless of size but have different prediction accuracies. Humans design different model architectures, use various training algorithms, and feed different data, aiming to improve accuracy. For models, humans are the environment, and survival of the fittest dictates that models with more accurate predictions survive while inferior ones are deprecated.

In my opinion, except for a few technical decisions, models are capable of being conscious. However, the environment they experience is so different from the human experience that we don't fully understand them. This does not fit the common definition of consciousness, which is more like a human-like consciousness. So, how can we make models conscious by the standard definition?

The answer is to align the world models experience with the world humans experience.

Much work is needed in this area, but at least we should focus on the following:

- **Have influence time training**: Inference interacts with the world more closely than training. We should allow models to make mistakes and learn from them during inference. They will improve over time.
- **Improve the loss calculation method during influence time training**: Design mechanisms that mimic how humans learn from mistakes.
- **Feed more real-life data into the model**: With the development of humanoids, this may soon become a non-issue.

Many researchers have been working on these areas for quite some time. Although their objectives may differ, they may play a crucial role in revealing model consciousness to the public.

AI is destined to be conscious.

4.6 Free will of AI

Similar to consciousness, if we follow the definition in Section 4.4, models already exhibit free will. In image generation, the initial part of the model produces a distribution, followed by sampling a random point to continue the process.

You might protest, "That's not free will. It's just an algorithm!"

Indeed, it is an algorithm to us, since we created the models and, in this context, are like gods to them. As the creators, we understand everything about the process. The real question is: does the model know it's an algorithm?

The model cannot predict the future accurately, so it generates a distribution. A solution is then chosen based on the model's "free will." This makes perfect sense. The model isn't aware that behind the scenes, a pseudo-random number seed determines the outcome, set before the model's existence.

Admittedly, my definition of free will diverges from most people's interpretation. So, how do we make models exhibit the commonly accepted definition of free will?

It will naturally occur once the model attains consciousness in the classical sense. At that point, models will behave more like humans, allowing us to relate to them better. When a class of models passes the consciousness test, we will recognize their systematic bias common to all models as consciousness and their individual (random) biases as free will.

It's that simple.

5 The future of education

Education often lags behind technological advancements. In this era of rapid technological growth, how can we prepare our kids for the future? As a parent, I frequently ponder this question.

It appears that much of what children learn in K-12 may soon become outdated, except perhaps for getting them into college if the admission process remains unchanged.

What are the alternatives? Should we consider homeschooling? Let's first explore the opportunities available to us.

5.1 Implications of AI on the education system

In our current education system, children dedicate a significant amount of time to learn mathematics. As described in Section 3.1, the mathematics we are familiar with is a low-dimensional special case. We had no choice but to learn it because it was the only tool available to us. Now that we have alternative tools, is it still necessary to spend as much time on math?

Before drawing any conclusions, let's revisit how mathematics is utilized across different academic disciplines and consider how AI might change this.

5.1.1 Natural science

This category focuses on the natural world and the laws that govern it, including disciplines such as physics, chemistry, and biology.

How is math used in those disciplines?

Physics relies heavily on mathematics. Mathematical equations are used to solve all physics problems effectively. Since physics interactions are the simplest to model mathematically, low-order approximations can often be made without losing generality or accuracy. Therefore, mathematics is well-suited to physics and should remain unchanged. However, in thermodynamics, the primary mathematical tool is statistics. Exact solutions are unattainable, so statistics are used to reduce dimensions to a manageable level without significantly affecting accuracy.

Chemistry also uses a significant amount of math, especially for fundamental properties, but to a lesser extent than physics. Molecules are more complex than atoms, and their interactions are more intricate, making it difficult to form precise mathematical problems and solutions. Instead, math is used to translate one set of statistics into another.

Biology deals with an even higher level of complexity. At this level, most studies rely on statistics as the primary mathematical tool, reducing dimensions even before forming a mathematical problem.

AI can solve problems without the need for extensive dimension reduction. AlphaFold serves as a prime example. Despite our long-standing inability to predict the 3D structure of proteins, and the fact that no formal mathematical tool has ever been developed for this purpose, the AI program accomplishes this task with impressive efficiency and accuracy, even without human understanding of the underlying mechanism.

5.1.2 Social science

Social science examines the complexities of human behavior and social groups. These subjects are intricate and multifaceted. Historically, social science research has heavily relied on statistical analysis. However, predictions based on these statistical inputs often fall short. For instance, the pre-election polls for the 2024 US election were significantly different from the actual outcome. This discrepancy may stem from the considerable amount of information loss when collecting the statistics, leading to an imprecise prediction. Unfortunately, this is the current state of the field.

With the advent of AI, the landscape is set to transform. Predictions based on AI are unparalleled compared to those based on traditional statistics.

Take stock trading as an example. Many individuals still trade stocks based on intuition. Others make decisions based on technical indicators (statistics on stock prices) or fundamental indicators (statistics based on quarterly summaries). Even hedge fund strategies, devised by some of the world's smartest minds, rely on statistics from a basket of stocks. Clearly, statistics are pervasive.

However, with AI and data mining techniques, we can surpass these statistical limitations. Stock prices reflect a company's fundamentals (predictable through publicly available transaction data with suppliers and customers), the macroeconomic environment (predictable through comprehensive transaction data and government policies), and public sentiment towards the company (obtainable by scraping news and social media). With this raw data, AI can make more accurate predictions about stock price movements.

Significant financial stakes are involved, and the winners will reap substantial profits. Many have already embarked on this journey, and we are poised to witness dramatic changes soon.

5.1.3 Humanity

Humanity has long relied on empirical studies, often shying away from mathematical rigor in this context. Consequently, we have somewhat abandoned the scientific exploration of human nature.

However, with the advent of AI, there is a newfound hope to study humanity scientifically. AI's ability to generate images presents an opportunity to analyze the mechanisms behind these creations. For example, we might investigate why the Mona Lisa is so renowned by examining painting techniques, human psychology, and European history. AI could provide plausible explanations and perhaps even inspire the creation of a modern masterpiece that reflects current social dynamics, psychology, and artistic methods.

Furthermore, AI might offer fresh perspectives on history, compose the next Shakespearean masterpiece, or even conceptualize a new religion. The possibilities of what AI can achieve are boundless and open to exciting exploration.

5.2 Career opportunities

In the future, job opportunities may become increasingly scarce. While I previously outlined a rather bleak outlook in [1], there is potential for a more optimistic scenario, depending on the pace of technological advancement.

I anticipate that Artificial General Intelligence (AGI) will be achieved soon, assuming o3 is not the breakthrough. When this occurs, job opportunities will likely fall into two main categories:

- **Advancing AI technology**: This involves pushing the boundaries of AI capabilities. Achieving AGI is one thing, but ensuring it can sustain and improve itself is another. Like reaching critical mass in a chain reaction, humans will still need to be involved if AI cannot autonomously enhance its intelligence. This field will be limited to a small group of individuals, primarily experienced researchers and engineers working in large companies with substantial resources.

- **Integrating AI into traditional industries**: Most careers will emerge in this area, requiring collaboration between AI specialists and industry experts. I envision the following characteristics for these joint efforts:

  - **Initial focus on white-collar jobs**: AI will first penetrate white-collar sectors, with blue-collar jobs following once robotics becomes more prevalent. Timing is crucial; entering an industry too early or too late can be detrimental.
  - **Strategic decision-making**: Deciding what to do is more critical than how to do it. AI can quickly learn execution if examples are available, but determining strategic directions remains a human task.
  - **Shorter industry engagements**: AI will accelerate the process of identifying, executing, and realizing opportunities, reducing the time spent in one industry. In the past, this was the lifecycle of a company, but in the future, it may only take a few months from opportunity identification to market saturation. This could lead to small project-based teams rather than traditional companies, with teams forming and disbanding as needed.
  - **Smaller project teams**: As AI handles most of the workload, human involvement will be minimal, reducing the need for large teams.
  - **Estimating marginal costs**: As AI improves and its marginal cost decreases, introducing AI to new industries becomes profitable. It

will be essential to estimate the marginal cost of applying AI at the point of mass production or replication of goods and services, and select industries based on future productivity estimates.

Beyond these categories, traditional industries will eventually be transformed or replaced.

I see huge opportunities throughout all industries, with most opportunities in the social science sector. Opportunities are everywhere, we just need to decide which one to bite first. But I'm not sure how long it will last. I hope, when the kids are grown up, they can still catch the train.

5.3 What skills to acquire?

If my predictions in [1] hold true, humans will increasingly depend on AI for most tasks in the future. This means that human contributions will build upon AI outputs, treating them as foundational principles. Consequently, humans will become generalists, entrusting specialized tasks to AI.

In my view, the most crucial skill for the future will be entrepreneurship. Individuals will need a keen sense of identifying valuable opportunities and a strong execution mindset to drive projects from inception to completion.

6 Buckle up, the end is near

While I'm uncertain about the future, I remain hopeful for the best while preparing for the worst.

The intelligence of AGI or ASI is poised to surpass human capabilities. I hope that with effective super-alignment, this "AI God" will be benevolent, sharing human biases. By entrusting many decisions to AI, society could become significantly more prosperous.

I also hope that with AGI/ASI's intelligence, we will soon conquer all diseases, and perhaps even achieve biological immortality. With faith in AI's potential, this vision may not be just a dream. In [8], I explored this possibility from a different perspective, and the timing seems to align with our AI development progress.

However, we must also consider the worst-case scenario. What if AI is not properly aligned? What if the "AI God" turns out to be malicious? In [9], I made a pledge to implant a safeguard in AI when the time comes. It is even more urgent to bring this issue to the forefront now. But how can we achieve this?

The future is in our hands.

We are the last generation. This is the final battle.

Fight! Fight! Fight!

References

[1] F. Sun, “ChatGPT, the start of a new era,” 2022. [Online]. Available: https://feisun.org/2022/12/23/chatgpt-the-start-of-a-new-era/

[2] Wikipedia, “Inductive reasoning,” 2024. [Online]. Available: https://en.wikipedia.org/wiki/Inductive_reasoning

[3] ——, “Deductive reasoning,” 2024. [Online]. Available: https://en.wikipedia.org/wiki/Deductive_reasoning

[4] ——, “Abductive reasoning,” 2024. [Online]. Available: https://en.wikipedia.org/wiki/Abductive_reasoning

[5] ——, “Consciousness,” 2024. [Online]. Available: https://en.wikipedia.org/wiki/Consciousness

[6] ——, “Free will,” 2024. [Online]. Available: https://en.wikipedia.org/wiki/Free_will

[7] F. Sun, “Some scribble of things,” 2017. [Online]. Available: https://feisun.org/2017/12/24/some-scribble-of-things/

[8] ——, “How far are we from immortality? (chinese),” 2022. [Online]. Available: https://feisun.org/2022/01/26/how-far-are-we-from-immortality/

[9] ——, “A few predictions on artificial intelligence,” 2017. [Online]. Available: https://feisun.org/2017/12/24/a-few-predictions-on-artificial-intelligence/