Journal of Human Cognition
2022, VOL. 6, NO. 1, 47-58
DOI: 10.47297/wspjhcWSP2515-469905.20220601

Minimal Turing Test and Children's Education

Duan Zhang[1], Xiaoan Wu[2], Jijun He[1]

1 College of Elementary Education, Capital Normal University, Beijing 100048, P.R.China

2 School of Marxism, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, P.R.China

ABSTRACT

Considerable evidence proves that causal learning and causal understanding greatly enhance our ability to manipulate the physical world and are major factors that distinguish humans from other primates. How do we enable unintelligent robots to think causally, answer the questions raised with "why" and even understand the meaning of such questions? The solution is one of the keys to realizing artificial intelligence. Judea Pearl believes that to achieve human-like intelligence, researchers must start by imitating the intelligence of children, so he proposed a "causal inference engine" to help future artificial intelligence make causal inference, pass the Minimal Turing Test, and even become a moral subject who can discern good from evil. This study attempts to provide some insights into the development of children's education from basic assumptions and construction goals of artificial intelligence, and to reflect on the causal model of artificial intelligence through children's education.

KEYWORDS

Causal inference engine; Counterfactuals; Free will; Science education

I Introduction

Artificial intelligence (AI) is defined as "the intelligence of a machine that enables it to conduct some human activities and attain complex goals". A decision or judgment made in the human brain is the result of billions of neurons interacting with each other, and the interaction between neurons is incredibly sophisticated and fast. Since technologies are not so advanced that we can understand how these billions of neurons interact, what contemporary artificial intelligence can do and has been doing is still only a simulation of our macroscopic cognitive model.

Alan Turing, one of AI pioneers, asked the question: How can a computer be said to be thinking like a human being? [1] He proposed the "Turing Test", in which a computer is asked a question, and if its answer does not allow one to distinguish it from a human being, then the computer can be considered a thinking machine. But it is difficult to imitate the intelligence of an adult, whose level of intelligence develops and advances all the time with his or her education and experience. So Turing argues that it may be better to start by imitating the intelligence of a child step by step before an ultimate realization of AI.

"Instead of trying to write a program that simulates the adult mind, why not try to produce one that simulates the children's? Then make it experience a proper educational process, and eventually it would have the brain of an adult. A child's brain is like a notepad from a stationery shop, with few mechanisms and many blank pages." [1]

In Turing's time, a child's brain was thought to be "like a notepad with few mechanisms and many blank pages". As J. Piaget pointed out, pre-schoolers had no causal thinking. [2] But psychological

research in the last three or four decades has largely disproved this conclusion, arguing instead that children's brains are rich in representations and learning mechanisms, which is why they learn so quickly and efficiently. For example, in causality, even children as young as 2 years old can make causal predictions, provide causal explanations, and understand counterfactual causal assertions[3-4] However, Turing's advocation of creating a robot with childlike intelligence before creating a robot with humanlike intelligence is supported by AI scientists.

Human intelligence manifests itself in many ways, for example, humans have vision, natural language, and reasoning ability. In Judea Pearl's view, the key step to creating a robot with the intelligence of a child is to master the causal relationship. This is because the understanding of cause and effect is one of the most universal and essential aspects of human cognition. [5] Only robots that have been built to understand cause and effect can be taught what they know about the world, for example, by telling a robot that physically cooling a patient with coronavirus disease with ice will not alleviate his or her breathing difficulties, or that if in reality a person sets a fire that causes a fire, teach the robot what would have happened if the arsonist had chosen not to set the fire. In practical and theoretical terms, there are at least two challenges to be addressed in creating a robot that understands cause and effect. Firstly, how does a machine acquire causal knowledge from its environment? Secondly, how does the machine process causal information? The first question is not of concern to Pearl, although it is an equally difficult one, involving a complex combination of the following inputs: active experimentation, passive observation, and input from the programmer. There are a number of reasons why this is not a concern. Firstly, if the problem of representation is not solved, it is not known how to store knowledge for future recall even if knowledge is acquired. Secondly, good representation is also instructive in how to acquire knowledge. And "one of the major contributions of AI cognitive research is the establishment of a paradigm 'representation first, acquisition second'. Answering the second question is his main contribution and effort, namely, proposing the way for machines to represent causal knowledge, and showing how this representation enables machines to access relevant information quickly and answer causal questions correctly. Pearl calls this the "Minimal Turing Test". How to pass the Minimal Turing Test has been a major effort of Pearl's for decades.

2 Blueprint for causal inference engine

How can a robot be made to enjoy causal conversations and answer causal questions with ease? Pearl [6] proposed the idea of equipping a robot with a causal inference engine (CIE) which would enable the robot to acquire and store causal knowledge and use it as a basis for causal inference and effective causal communication. The CIE has the following goals: 1) the representation and storage of causal knowledge; 2) the simulation of how humans make causal inferences; 3) the implementation of knowledge iteration and update; 4) the implementation of the entire process algorithm. Figure 1 below is a detailed blueprint for CIE proposed by Pearl.

This blueprint also reflects how causal models work and how they interact with data in today's scientific applications. Next, the theoretical assumptions will be described and illustrated by exemplifying the relationship between smoking and lung cancer.

In Pearl's view, every causal inference task necessarily relies on assertions that go beyond the data and are derived from knowledge (K), i. e., "past observations, past behaviors, education and cultural practices, and all others relevant to the target problem". From a hermeneutic perspective, knowledge can be understood as "foresight", "at a micro level, it is related to genetic factor, family background, childhood experience, learning and working process, and especially traumatic experience different from others", and "at a macro level, it is linked with cultural and social

backgrounds, traditional customs and concepts, national psychological structures and preferences, and cognitive level and ethical tendency in a specific era." [7] Broadly speaking, the sources of knowledge can be divided into three parts: evolution, cultural transmission, and individual experiences, all of which construct people's understanding of the world and in doing so, solve problems and accomplish tasks.

In Figure 1, the "knowledge" is placed in the dashed box because it is still hidden in the head of the reasoning subject. But when specific problems are examined, the stored knowledge is called upon, at first perhaps as a rough and macroscopic set of assumptions (A), some patterns which, if known adequately, will also be further refined into an explicit equation or logical expression. These assumptions are made by the researcher for scientific reasons or firm beliefs to defend. For example, when studying whether smoking causes lung cancer, a researcher in this field should have an understanding of the relationship between smoking, lung cancer, tar and genes.

In Pearl's view, the way in which these patterns and assumptions are represented is important because there are three aspects involved. Firstly, it affects the correct and precise way in which one states these assumptions. If the language used is not concise and the presentation is not intuitive, it does not facilitate but hinders the perceived knowledge of the assumptions. Secondly, it affects the derivation of conclusions from these assumptions, which naturally affects willingness and confidence in their use if the process of derivation is too cumbersome. Thirdly, it is easy to extend or modify these assumptions when more convincing evidence is available. Consider the inefficiency and complexity of the symbolism school of AI in recording the spatio-temporal information structure of 4D objects in the formal language of 1D modern logic, and the frame problem faced in expressing the dynamic laws of 4D objects. [8] Graph language is a very intuitive and readable language with irreplaceable advantages for perceiving, understanding and representing knowledge of the world. The graph theory has also made great strides in recent decades.

The graph language qualitatively encodes the researcher's causal assumptions about the situation under study and is presented in such a transparent way that the causal assumptions embedded in them can be read easily and quickly by anyone, with the implications being discussed and debated by experts in the field to ultimately confirm or deny the validity of the diagram itself. Without a causal graph to clearly express the substantive causal assumptions, there is a risk of error in making inferences. D. Rubin gives an example of how industry legend R. A. Fisher made the fundamental mistake in making an adjustment for a mediation outcome when estimating a treatment effect. [9]

"The researcher must draw up and present a causal graph that reflects his qualitative judgment of the topology of the causal processes involved in a research topic, or more ideally, reflects the

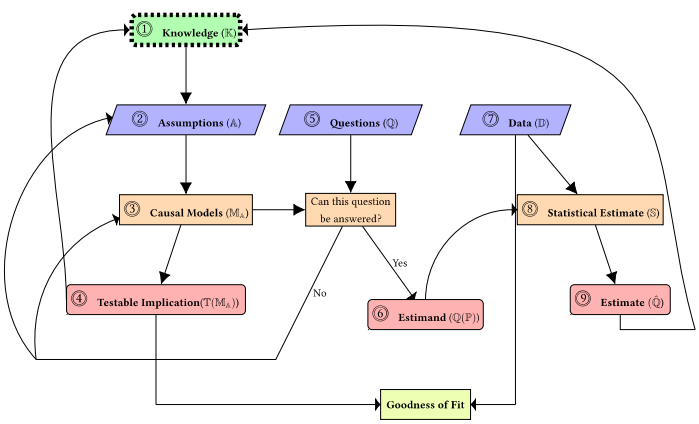

Fig.1 Blueprint for Causal Inference Engine of Artificial Intelligence in the Future

consensus of researchers on that research topic in the field ([10], p.67). "

The graph language is of course only a theoretical resource for Pearl's structural causal models, and when it comes to dealing with counterfactuals, the graphs as a tool are not enough, so he brings in structural equation models. A structural causal model *M* is an ordered triple <*U, V, F*>, where *U* denotes the set of extrinsic variables, representing those factors that are not considered by the model. To determine a nomologically sufficient condition of an event, it is required, according to the theory of relativity, to describe some cross-section of the entire inverse time light cone that precedes the event. The number of factors is unimaginable, even if one were to granulate it relatively coarsely, and some trade-offs must be made, so *U* is used to generalize those factors that are left out for the moment. *V is the set of intrinsic variables, i.e., those that are determined by extrinsic variables or other intrinsic variables. F is the set of structural equations. For each intrinsic variable, there is a corresponding structural equation, i.e., for each intrinsic variable* $V_i \in V$ *there is a mapping fi from* $U_i \cup PA_i$ *to* V_i*:*

$$v_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n \tag{1}$$

where $U_i \subseteq U$, $PA_i \subseteq V/V_i$, and *PAi* is generally referred to as the parent set of V_i. The variables in *PAi* have a direct causal effect on *Vi*, and the choices of *PAi* are not arbitrary, but "express the modeller's understanding of those variables that nature (or the Creator) will necessarily explore before deciding on the value of V_i".

First, each causal model has a causal graph (denoted *G*(*M*)) corresponding to it. Each node in this graph corresponds to an intrinsic or extrinsic variable. If a variable is on the right-hand side of a structural equation, an arrow is drawn from this variable to the variable on the left-hand side of the equation, and eventually a causal graph is obtained. If the causal graph is a directed acyclic graph, the corresponding model is called a semi-Markovian model, which is characterized by that the values of all the variables in the model are uniquely determined if the value of *U* is determined to be *u*, or a distribution defined over the set of intrinsic variables *V* is uniquely determined if a probability distribution of the variables in *U* is determined.

If the causal graph is a directed acyclic graph and the extrinsic variables U are also jointly independent, the model is called a Markovian model. In many cases, it is certainly very difficult to determine the exact structural equations, especially if they are non-linear. However, even if only the causal graph representing the qualitative relationship between the variables is known, many causal questions can be answered, i. e., it is possible to determine algorithmically whether the above graphical model is adequate for the identification of the required causal effects and which covariates need to be measured.

Secondly, each structural equation（$v_i=f_i(pa_i, u_i)$）characterizes a stable and autonomous physical mechanism, i.e., it can be conceived to change it without changing other mechanisms. Essentially, knowledge is organized in a modular manner so that the intervention effects can be predicted with very little additional information. Again, a strong foresight of the modeler is implied in the construction of the above model: how the variables causally affect each other. For example, identify the set (PA_i) of all variables in *V* that have a direct causal effect on V_i. The benefit of this causal information to the model is that the causal relationship is a cornerstone of human knowledge construction, and judgments based on it are more meaningful, acceptable and reliable. Based on the above description, the model depicts a specific situation in which, when an intervention occurs for one (or some) of the variables in the model, the intervention is local to the model and does not affect mechanisms other than those associated with the intervening variable, according to autonomy.

Before construction of a causal graph, it is necessary to well define the causal relationship. Philosophers have debated over the nature of cause and effect for thousands of years, but have never reached a consensus. Compared to the painstaking efforts of philosophers [11], [12], Pearl's

establishment of the causal relationship between two variables is a bit of a "children's play".

"If variable Y 'listens to' variable X and determines its own value based on what it 'hears', variable X is a cause of variable Y([10], p.XX). "

First, "listen" relationship is a weak one. To establish such a relationship, the relationship between variables is not necessarily causal, but can be logical or purely mathematical. For example, *WF* denotes final weight, *WI* denotes initial weight, and *Y=WF-WI* denotes weight gained. When drawing the causal graph, it is still possible to draw arrows pointing from *WF* and *WI* to Y. Second, one may feel that this suggestion seems too hasty, but there is actually no need to worry, because the conditional independence between variables or counterfactual variables implied in this graph, i.e., testable implication (T(MA)), can be identified by using the created causal model or causal graph ( ) and using techniques such as d- separation and bipartite network.

And if the data collected contradicts its implication, the graph has to be corrected. Again, given that it is almost impossible for the data and the graphical model to agree perfectly, another engine is needed to calculate the "goodness of fit", i.e., the degree to which the model is compatible with the data. With the data, it is possible to challenge the validity of certain causal assumptions and to determine in what sense an alternative causal relationship would better explain the observed data. In the aforementioned study of smoking-lung cancer, the qualitative assumptions accepted by the researcher can be characterized as the following causal graph and set of structural equations with unknown parameters, as shown in Figure 2.

According to the graph theory, those substantive causal assumptions encoded in the graph are expressed by negative causal assertions, i.e., missing causal links between nodes in the graph. There are two types of missing links [3]: the first type is the missing arrow, for example, the absence of arrows between variables *X* and *Y* in Figure 2 that point directly to each other. It indicates the following presupposed assumptions: Factors that are ignored as directly affecting *X* are unrelated to those that directly affect *Y*, and *X* has no causal effect on *Y* when the value of *Y*'s parent variable (i.e., variable *Z*) is fixed by intervention. In the modeller's view, there is no direct causal effect of smoking on lung cancer and the effect between the two is mediated by the accumulation of tar in the lung. The missing arrows therefore encode exclusion restrictions, i.e., adding variables to the equation that are excluded from the equation (e.g., adding the variable *X* to the equation *z=f2* (*x*, *u2*) does not change the outcome of the equation. In potential outcome model notation this is expressed as follows:

$$Y_z = Y_{zx}, Z_x = Z_{xy}, X_y = X_{zy} = X_z = X \tag{2}$$

The second type is the missing bidirectional dashed arcs. For example, in Figure 2, there is no bidirectional arc linking the variables *X* and *Z*. It indicates that factors directly affecting *X* are unrelated to those directly affecting *Z*. That is, in the modeller's view, factors that affect smoking are unrelated to those that affect tar accumulation in the lung other than *X*. Specifically, the missing bidirectional arcs encode independence restrictions. For example, the absence of bidirectional arcs between *X* and *Z* (and between *Z* and *Y*) means that U_1 and U_2 are independent in $P(u)$, the value of *y* is entirely determined by u_1 if the value of *z* in $y=f_3(z, u_1)$ is determined, and the value of *z* is entirely determined by u_2 if the value of *x* in $z=f_2(x, u_2)$ is determined. Since U_1 and U_2 are independent, it is natural that Z_x and *Yz* are independent, and by the same token that Z_x and *X* are independent.

So the missing arc in Figure 2, rather than the existing one, should have received more attention and more careful validation. After all, adding an additional bidirectional arc causes at worst an unrecognizable parameter, but deleting an existing bidirectional arc may lead to erroneous conclusions. In modeling, the existence of a bidirectional arc between any two nodes should be assumed by default, and a bidirectional arc should be considered non-existent only if this can be well justified, e.g., it is impossible for two variables to have a common cause but a selection bias to exist.

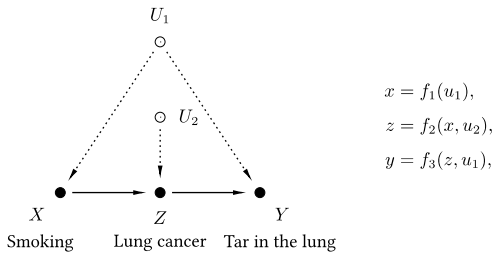

Fig.2 Causal graph of smoking impact on lung cancer

Although it is never possible to know all the factors that affect a variable, an individual can sometimes conclude based on his/her substantive knowledge that the influence of a possible common factor is not that important.

Next, the causal language is used to express the causal question ($\mathbb{Q}$) which can be an intervention or prospective counterfactual question, a retrospective counterfactual question, etc. As in the example in Figure 2, the causal effect of smoking on lung cancer is required and can be formalized in causal language as $P(y|do(x))$ or, assuming that Tom has smoked for 30 years and has lung cancer, the probability of lung cancer for him if he did not smoke can be formalized in causal terms as $P(yx|x', y')$.

Then, given the causal model and the data, can the question above be answered? Pearl gives the graph-based $_{zu}$transformation rules for prospective counterfactuals, ([13], pp. 85-86) gives a three-step procedure for calculating reflexive counterfactuals, ([13], p. 206) gives mediation formula calculating the Natural Direct Effect (NDE) and Natural Indirect Effect (NIE), and ([13], p.132) even gives a definition of an algorithmically realizable causal model regarding the actual causal questions under determinism. [14] Whether Pearl's framework still leaves out some counterfactuals and causal questions is debatable ([15], p. 1136) but there is no doubt that most causal questions and concepts can be characterized and answered within the framework of the causal model.

If the above question is not answerable, i.e., the answer is "No", it is required to re-examine the assumptions and the causal model for soundness or further optimize the model. If the answer is "Yes", based on the formula and calculation rules given above, an Estimand (Q(P)) can be obtained.

The duty of the estimand is to tell us how to convert the statistic into an expression based on the assumptions of the model, and this expression is logically equivalent to that of the causal question (e.g., $P(y|do(x))$) ([10], p.XXI).

In the smoking - lung cancer example illustrated in Figure 2, it is required to first consider the factors affecting tar accumulation in the lung, then consider the effect of tar accumulation on lung cancer, and finally do a weighted calculation:

$$P(y|do(x)) = \sum_z P(y|do(z))\boxtimes P(z|do(x)) = \sum_z P(y|do(z))\boxtimes P(z|x)$$
$$= \sum_z \sum_{x'} P(y|z,x')\boxtimes P(x')\boxtimes P(z|x) = \sum_z P(z|x)\boxtimes \sum_{x'} P(y|z,x')\boxtimes P(x') \quad (1)$$

The final formula on the right side of the equation is the 'estimand', a way of calculating the target quantity of the problem, which is expressed as a probability, i.e., an estimate of the target quantity can be obtained from observable data (D). The estimand may not be unique. Quite a lot of estimands may be obtained with Do-Calculus, yet some of which are not observable of their variable. For example, the following equation:

$$P(y|do(x)) = \sum P(z|x) \sum P(y|z,u)P(u) \quad (2)$$

The data for gene (U) in the above equation are not observable. The algorithmization of counterfactuals is also possible if the functional relationship between a variable and its parent set is stated exactly in the causal model. For example, in Figure 2, Tom smokes and gets lung cancer, the

probabilityof not getting lung cancer if hedid not smoke is: $P(yx|x', y')$.

Based on the three steps of the counterfactual algorithm ([13], p. 206), the estimand and the estimate of this counterfactual can be calculated.

With the estimand and the data, an estimate（Q）is obtained. However, the estimate obtained is only likely to be an approximation as the data collected is always finite and it belongs to a theoretically infinite aggregate, leaving the problem of selection bias. But with powerful machine learning techniques, statistics has developed ways to cope with this uncertainty and to give a statistical estimate (S) of the degree of uncertainty, such as maximum likelihood estimation, propensity score, confidence interval, and significance test ([10], p.XXII). If the model is correct and the data are sufficient, one answer to the causal question is also obtained. If it is true, for example, that smoking causes lung cancer, it can be added to one's own scientific knowledge base, thus achieving growth of knowledge.

In short, what Pearl seeks to achieve is one aspect of the human capacity for consciousness: the ability to make causal inferences. In his view, vast amounts of causal information, whether inherited from genes or acquired from experience, is stored in the human brain, and this knowledge can be described in a concise mathematical language: logic, graph, mathematical equation, counterfactual variable, etc. For a new problem, the stored knowledge is called upon to solve the problem in combination with acquired information and data. The practice itself also allows us to gain new knowledge or realize our own powerlessness. Undoubtedly, intelligence does not happen overnight. From weak to strong intelligence, it is a process of learning knowledge, probing unknowns, solving problems, designing new experimental methods, and exploring the Nature. In this process, intelligence itself undergoes a process of constant renewal and iteration and becomes more and more visible. Perhaps, it will pass the Turing Test one day.

3 Counterfactual algorithm and illusion of free will

For all efforts around CIE, the "algorithmization of counterfactuals" is one of the key results of Pearl's work given the importance of counterfactual inference for the realization of intelligence.

"Counterfactuals are the cornerstone of ethical behavior and scientific thinking. The ability to retrace past actions and imagine other possible scenarios is the basis of free will and social responsibility. The algorithmization of counterfactuals makes it possible for "thinking machines" to acquire this ability unique to our human being and to master the way of thinking about the world that is still unique to us ([10], p.XVII). "

"Will" can be understood as the will to do something, and "free will" means that your decision to do something is entirely up to you. For example, your decision to smoke or not is entirely made by you based on your rational and deliberate evaluation. The above description of free will is obviously very crude. If analyzed rigorously, it can be revealed that any choice is determined and that there is no such thing as true free will. Description can be made at two levels: At the neuronal level, all processes are deterministic (for neuroscientists, the human body and brain are automatic systems that follow deterministic laws), and at the cognitive level, we truely feel the autonomy of choice ([10], p. 337). It seems that we, born as human beings, have always possessed a sense or illusion of free will, like "I am the master of my fate, I am the captain of my soul". Many of our behavioral choices come entirely from us, for example, if I want to reach for this bottle, I can reach for it, and if I don't want to, I will not reach for it.

In which sense, does the machine have "free will"? Pearl points out, it means that the machine has free will "if the machine has the intention to do $X=x$, and then chooses to do $X=x$ after realizing this intention". What is the purpose or function of the illusion of free will that evolution has gone to

the trouble of giving us? This is the question of function. If the "illusion" of free will has a legitimate function, how does one give free will to a robot that is determined by the algorithm? This is a question of simulation. The importance of functional issues has to be established before the project can be motivated and driven, even if it is extremely difficult to implement the project.

First of all, Pearl points out that it is because of free will that it is possible to speak meaningfully about intention and to make intention comply with reason by envisaging other possible alternative situations, i. e., by giving a rational interpretation to the action itself. For example, when a person considers buying Huawei (*do*(*A*)), Xiaomi (*do*(*A'*)) or Apple (*do*(*A**)), he may examine the preferences for the different consequences of the actions themselves, including vanity, performance required, affordability, etc. He is actually determining the utility function, and after considering these possibilities and their implications, finally decides to buy Huawei, as shown in Figure 3.

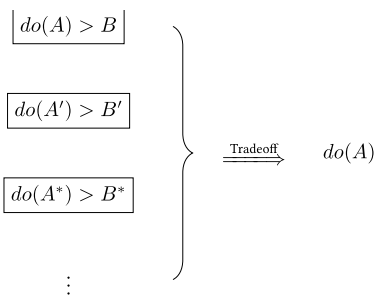

Fig.3 Tradeoff of All Possible Actions

The rational thinking process described above has been algorithmized within the framework of a structural causal model. Further, once a robot has free will, the uncertainty of its communication with us can be reduced because it is able to observe its own intention and take another action in a specific situation once the machine believes it has free will.

For example, your household robot turns on the cleaner for cleaning in the early hours of the morning while you are asleep, you are woken up and are very annoyed, so you tell it loudly: "You should not use the cleaner while I'm asleep." With this counterfactual expression, we are asking the robot to go back in time to re-experience its own experience and to revise the procedures that determine its behavior. For example, the robot understands through this counterfactual communication that you were angry because you were woken up, and its built-in schema enables it to know that too much noise can affect sleep, and that the noise was generated by the cleaner. With the counterfactual algorithm, it can counterfactually think: what would have happened if the cleaner had not been turned on? It can infer that no noise will be produced, and a sleeping person will not be woken up, and will not get angry for being woken up. Since your anger makes it aware of a possible option to avoid making you angry again, the next time it will choose not to turn on the cleaner while you are asleep. In fact, we often communicate with our children in such a way that we can slowly discipline their behavior and lead them to the right path. (Figure 4)

Therefore, if the robot has the illusion of free will, there is an intention behind each of its actions. By pointing out the inappropriateness of its actions, we can make the robot to think about its intentions, reflect on its mistakes, envisage the possible outcomes of alternative actions, and then adjust its behavioral choices according to its utility and preferences for the outcomes, further optimizing its actions. And only when the robot is able to make other choices, or to think counterfactually, can we effectively discuss the counterfactual concepts including credit, blame, responsibility and regret. Pearl is certain about the realization of a strong AI with causal understanding and agent capabilities, based on the solution of his three levels of questions about the ladder of causation. What is to be awaited is merely the implementation and realization by

engineering means.

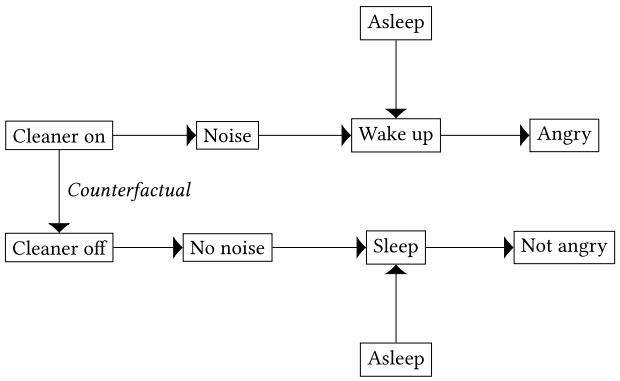

Fig.4 Robot's Counterfactual Inference Process

"The ability to reason about one's own beliefs, intentions and desires has been a major challenge for researchers in the field of AI, and has defined the concept of "agent". Can machines be conscious? What distinguishes software agents from ordinary programs? Such questions have plagued generations of elites ([10], p.339). "

Pearl does not think he has fully answered this question. But he believes that his algorithmization of counterfactuals is a major step forward in understanding these questions and in making consciousness and agents a computable reality.

Finally, in his view, if a software package is used to enable a thinking machine with the capabilities of an agent, the package should consist of at least three components: "a model of the world, a model of the software itself, even if it is only an overview of the software blueprint, and a memory to record how intentions in the mind correspond to events in the external world" ([10], p. 340). With the data recorded in memory, it is possible to trace the decision-making process to explain and modify one's behavior. Pearl probably thinks that a robot with the CIE is then conscious, but whether the reasoning ability is equivalent to subjective consciousness remains an open question. Even if it is possible for a robot to act as an agent, is it a moral agent? P. Taylor identifies five abilities that are most important for a moral agent: "the ability to make judgments about right and wrong; the ability to engage in moral deliberation, that is, to consider and weigh the various moral reasons for and against alternative actions; the ability to make decisions based on these reasons; the ability to exercise the determination and willpower necessary to implement those decisions; and the ability to explain one's failure to implement those decisions." ([16], p. 14). In other words, to know right from wrong, one must reflect on one's actions, consider the various possible options, understand the consequences of the choices, and make final decisions by moral or social norms. Since the machine has the ability to make counterfactual inferences, and the above five abilities are implemented algorithmically, it seems to be viable as a moral agent.

4 AI CIE and children's education

Zhu Songchun points out that "those who engage in visual cognition must go back to the Stone Age to understand the functions of objects, and those who engage in language must go back to the origin of language." [17] Similarly, to realize AI, we have to start from children. The goal of general artificial intelligence is to create "agents with autonomous perception, cognition, decision-making, learning, action-taking and social collaboration capabilities that are consistent with human emotions, ethics and morality". [18] If such an agent is to be achieved as envisaged by AI designers by

slowly nurturing and teaching a child agent, i.e., by providing it with science education and moral education, it is similar to the education of children since there is no theoretical difference between creating an agent with "decision-making ability and moral-emotion" and raising a child with the same ability and emotion.

Currently, most discussions on science education and AI focus on the application of AI technologies to science education. In fact, there is a strong link between the children's approach to AI and children's education. Firstly, how to educate a child to be academically and morally competent may inspire us to build a moral agent. Secondly, the thinking of scientists in the process of building a general agent may also have implications for the education of children.

5 AI CIE to Children's Education

It can be found in the construction of the agent: Firstly, intellectual and moral education are in fact inseparable. In most of the children's education literature, science education and moral education are separated. However, in the view of AI scientists, the ability to make complex logical arguments and causal inferences is a goal shared by good science education and moral education. In Pearl's AI construction, counterfactual reasoning and causal inference, in addition to their role in scientific inquiry, are necessary conditions for making a moral agent capable of discerning good and evil possible. Whether in the judgment of right and wrong or in scientific inquiry, agents always reason and think on the basis of observed facts and in strict accordance with the causal and logical connections between events. The choices are driven by various decision functions and value functions through rational consideration of the outcomes and preferences of possible choices. This is consistent with the 'rational man' supposition in economics, namely, a person's actions and decisions are driven by interests and values to maximize their own interests. In short, both scientific reasoning and ethical consideration must be based on facts and follow reason. Actually, even counterfactual situations against reality are constructed based on reason and logic, rather than being determined by purely subjective ideas and unreasonable emotions.

Secondly, science education cannot be merely the inculcation of scientific facts, nor can moral education be merely a form of discipline training. Human thinking activities and behavioral decisions are driven by two major factors (models and expressions) [18], the *U*-system and the *V*-system, where the *U*-system "is a stable social regulatory force eventually formed in the process of long-term social practice, rule evolution, and multiple games". As people may sigh, "a person sometimes has to do what he or she was forced to do to be a player in the human world." The ethics and norms of society, as well as the objective laws of physics, are a constant "constraint" on the decisions of every "social being" in the world. The *V*-system is "a value system evolved by the common influences of the environment, biological properties and hormones in the evolutionary process". It is a collection of value functions of many dimensions, such as biological evolution, culture and religion, and the individual value. The bag of norms and virtues will swell as society evolves and is not constant but "a temporary state of quasi-equilibrium reached by groups of people in competition and cooperation, subject to external physical and causal constraints". However, the old equilibrium will be broken, and a new one established as the society evolves. It is highly impossible to encode these complex and sometimes conflicting *U* and *V* systems into the decisions of robots and to guide them in how to weight specific situations.

Isaac Asimov proposed the three laws of robotics in an attempt to foster moral agents. These three laws are consistent with the idea of expert systems based on rules, all of which have their inevitable problems. First, the laws themselves may not always be ironclad, and all trade-offs should still be made, as our experience reveals. Second, it is perfectly possible to conceive a robot that conforms to the three laws mentioned above, but still believe that it should not act, or that it should

act but violate one and two of the three laws. In fact, Asimov, in his science fiction "I, Robot" in which his three laws were proposed, has already pointed out that these laws would leave robots "not knowing what to do".

Asimov's thinking was top-down [19] as stated by W. Wallach and C. Allen, Pearl's was a hybrid top-down and bottom-up approach. Pearl encoded norms into the robot's program and incorporated them into his rational thinking: Once a machine is encoded with self-awareness, it will then have "empathy" and "a sense of fairness", because the computational principles or algorithmic processes are the same for these two emotions, except that another agent is added to the equation([10], pp.342-343).

Nevertheless, moral education is not exactly a form of discipline education, "moral education is not an education in doing what is right, but in the nature of moral thinking and in the skills and abilities to make decisions" [20], i.e., giving robots the illusion of free will and the ability of causal reasoning so that they can understand right and wrong and modify their behavior in the context of their interaction with us. Pearl argues that for the realization of a moral agent who knows right from wrong, we should start from training it with discipline. For the extensiveness of rules and the limitlessness of knowledge, it is ultimately necessary to empower robots with the abilities of thinking and discernment so that they can correct, evolve and improve themselves for adaptation to complex situations. The science education and moral education of children also follow this path.

6 Children's Education to AI CIE

First, as mentioned earlier, children have the ability to make causal inferences and know how to work with people at an early age. If a robot is to be programmed, it must have the ability to infer the thoughts and reactions of others. Otherwise, it will not be able to understand what is ethical and moral. A robot must first be able to imagine the consequences of actions and understand the criteria of good and bad to determine the goodness or badness of these consequences, and thus ultimately decide on its own actions before it can be rated moral and ethical.

Second, just as it is impossible to teach a child everything at once, it is also impossible to encode all knowledge about the world into an agent. The agent created should be able to acquire knowledge, to update and iterate on it. Fundamentally, science education for children is about making scientific concepts and principles, which are far from experience and common sense, understandable to children and enhancing their understanding of science and technology in order to stimulate their interest in the subject and their creative potential. Science education is not essentially "education in scientific knowledge", but in scientific thinking. For example, the key to teaching the periodic law of the elements is not to make pupils memorize the entire periodic table, but to make them understand the tortuous process of discovering these elements through his torical review, to reveal the ideas they contain, to help them understand the nature of chemical elements by making connections between atomic weights and other properties of elements. Comprehension itself is a matter of "understanding". For students whose logic is confused and whose thinking is completely out of order, even the most lively education will only achieve half the result. The skills of logical and cause-and-effect thinking seem as simple as a boxing routine, but achieving the "unity of man and boxing", that is, to be logical, clear and focused, requires long-term logical education and mathematical training. Computers are far superior to humans in many abilities because they rigorously implement algorithms in their actions.

Third, there are many descriptions of children's scientific practices in children's science education[21], and the level of intelligence demonstrated by children reveals that reasoning is only one aspect of intelligence which includes induction, analogy, description, association, abstraction, and systematization. To realize the formulation of questions, the construction of models,

observation and reflection, experimentation and creativity, and the ability to learn by example, breakthroughs are required in AI areas, such as computer vision. After all, 70% of the human cerebral cortex processes visual information. A further review of the causal inference engine reveals that the implementation of general AI still has a long way to go, and only a blueprint is envisaged in terms of causal inference. Moreover, the algorithmization of counterfactuals in the engine is actually quite different from our counterfactual inference since we do not need to determine functional relationships between variables when inferring counterfactuals.

Finally, even if a child is given the best and the most decent education, he or she may still end up to be a "bad" person who does not know good from evil. Even if a robot could be given the illusion of free will and the abilities of causal and counterfactual inference, there is no such a process of education that would absolutely and ultimately shape the robot into a moral agent knowing good and evil. After all, evil is always born unexpectedly as in the case of Skynet in the film Terminator, which is bent on ruling humanity and was clearly not designed to turn against its creator one day.

References

[1] Turing, A. Computing Machinery and Intelligence[J].Mind, 1950, 59(236): 433-460.

[2] Piaget,J.The Child's Conception of Physical Causality[M].New York: Harcourt, Brace, 1930.

[3] Hickling, A., Wellman, H. The Emergence of Children's Causal Explanations and Theories: Evidence from Everyday Conversation[J].Developmental Psychology, 2001, 37(5): 668-683.

[4] Gopnik,A.,et al. A Theory of Causal Learning in Children: Causal Maps and Bayes Nets[J]. Psychological Review, 2004, 111(1): 3-32.

[5] Mei Jianhua. Artificial Intelligence and Causal Inference-On the Singularity Problem[J]. Philosophical Researches, 2019,(6): 86-95.

[6] Pearl,J.The Seven Tools of Causal Inference, with Refections on Machine Learning[J]. Communications of the ACM, 2019, 62(3): 54-60.

[7] Zhang Jiang.A Re-Discussion of Imposed Interpretation[J].Social Sciences in China, 2021,(2): 4-23. 204.

[8] Ye Feng. On the Role of Language in Cognition[J].World Philosophy,2016,(5): 72-82;161.

[9] Rubin, D. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions[J]. Journal of the American Statistical Association,2005,100(469):322-331.

[10] Pear,J.Mackenzie,D.The Book of Why:The New Science of Cause and Effect [M].Translated by Jiang Sheng and Yu Hua. Beijing: China CITIC Press, 2019.

[11] Papineau,D.The Causal Structure of Reality[J].A Long (double-length) Version. Monist.forthcoming,2022,1-51.

[12] Lewis,D.Causation[J]. The Journal of Philosophy,1973,70(17):556-567.

[13] Pearl, J. Causality: Models, Reasoning, and Inference[M]. New York: Cambridge University Press, 2009. Halpern, J., Hitchcock,C.Graded Causation and Defaults[J]. The British Journal for the Philosophy of Science, 2015, 66(2): 413-457.

[14] Imbens,G.Potential Outcome and Directed Acyclic Graph Approaches to Causality:Relevance for Empirical Practice in Economics[J].Journal of Economic Literature,2020,58(4): 1129-1179.

[15] Taylor,P.Respect for Nature:A Theory of EnvironmentalEthics[M].Princeton: Princeton University Press,2011.

[16] Zhu Songchun.Artificial Intelligence:Current Situation,Tasks,Framework and Unification[J].The Vision Seeker,2017,1-68.

[17] Zhu Songchun. Three Readings of Chibi Fu, and Interpretation of the Balance between"Heart"and"Principle"from AI Perspective[OL]. https://mp.weixin.qq.com/s/MuOjSBeWcmmc1t9lQC6nsA.2022-01-06.

[18] Wallach,W.,Allen,C.Moral Machines: Teaching Robots Right From Wrong[M]. Oxford: Oxford University Press,2008.

[19] Hall, R. Davis, J. Moral Education in Theory and Practice [M]. Translated by Lu Youquan and Wei Xianchao.Hangzhou: Zhejiang Education Press, 2003.

[20] Li Yanbing. Science Education in Primary and Secondary Schools in Piaget's Early Scientific Works[J]. Research in Educational Development,2010,30(20): 68-72.