Is AI intelligent because of its instrumental rationality?

Thesis submitted for the examination of Subject 199 in Philosophy

Candidate Number: 1042151

Candidate School: Honour School of Mathematics and Philosophy (Part C)

Word count: 19111

Abstract

This thesis argues against the claim that AI is intelligent due to instrumental rationality, refuting both the reduction and emergence thesis. It contends that intelligence cannot be reduced to instrumental rationality and highlights issues in AI development and application. Instead, it proposes the motivation adaptation approach, where intelligence arises from network of generative motivations and the ability to adapt. This alternative is conceptually intuitive, avoids counterexamples, and provides clear development goals and foundations for ethical development. Thus, the thesis concludes that AI's intelligence is not due to instrumental rationality but rather its motivation adaptability.

Keywords: artificial intelligence, rationality, casual learning, emergence, ethics, adaptation

Contents

Abstract 2

1. Introduction 4

2. Defining rational agency approach 6

   2.1. Instrumental rationality in the field of AI 6

   2.2. Rational agency approach at work 8

   2.3. Specification of rationalities 10

3. Problems against strongly construed rational agency approach 15

   3.1. Problem of cognitive slack 15

   3.2. Problem of shallow success 18

   3.3. Problem of abstract robustness 20

   3.4. Problem of causal learning 24

   3.5. Problems in conception 28

4. Problems against weakly construed rational agency approach 30

   4.1. Problems in development 30

   4.2. Problems in application 34

5. An alternative approach: motivation adaptation approach 39

   5.1. Defining motivation adaptation approach 39

   5.2. Answering the problems against rational agency approach 44

6. Conclusion 51

Bibliography 52

1. Introduction

What do we want from AI? One natural answer is its usefulness. We dedicate a great deal of attention and resources to developing AI because we want AI to be useful. In philosophical terms, we express this notion of usefulness through the concept of *instrumental rationality*. To be instrumentally rational, one delivers good performance for ends. As instrumental rationality is a good thing, many AI theorists propose that we define AI as an instrumentally rational machine that can perform well for our ends. This view of AI is featured in Russell and Norvig's textbook *Artificial Intelligence: A Modern Approach,* which has been adopted for use by at least 1500 schools in 134 countries or regions by 2022.

However, we seldom believe good performance alone is sufficient to demonstrate intelligence. In many scenarios, good performance happens without intelligence. As I will demonstrate in the following thesis, it happens when students solve a mathematics problem by blindly following the taught procedure. It happens when someone earns a lot from rising house prices in his neighbourhood without hindsight. Moreover, since no machines now understand language, it seems to happen when Chatbot provides satisfying responses to your questions.

Thus, when AI theorists define AI as an instrumentally rational machine, concerns arise that AI might not be genuinely intelligent despite its high performance. As AI stands for "artificial intelligence," the notion of unintelligent AI poses a significant theoretical issue[1]. If AI theorists insist on maintaining that AI is an instrumentally rational machine, they must assert that its intelligence stems from its instrumental rationality. Tconundrumdrum gives reasons to investigate the thesis question: Is AI intelligent due to its instrumental rationality? Because if the answer is negative, it provides a compelling reason to reject the idea that AI is solely an instrumentally rational machine.

In reply to the thesis question, there are two kinds of affirmative responses. The first response is to argue that intelligence can be reduced to instrumental rationality. Following this response, a precise concept of instrumental rationality is required to support the reduction. When we find this precise notion of rationality and succesfully develop rational machines, they will, by

[1] This thesis will not discuss the disjunctivist's position that machine's intelligence is different from human intelligence for two reasons. First, disjunctivist has significance burden to explain what machine intellitence is. Second, as I will argue in section 4.2, many ethical applications of AI are founded on machine being intelligent in human sense.

definition, be intelligent because of their rationality. In comparison, the second response argues that intelligence can emerge from instrumentally rational machines. According to them, instrumental rationality is the ultimate pursuit in the field of AI, and intelligence is a by-product that emerged from this pursuit. Therefore, they argue that we should only focus on making instrumentally rational machines without explicitly positing intelligence as a goal. Then the development, if machines turn out to be intelligent, we will say that they are intelligent because of their instrumental rationality.

Nonetheless, neither kind of response works. In the following thesis, I will present my main arguments in four sections. In Section 2, I will explain precisely what instrumental rationality is and how it gives rise to the rational agency approach. In Section 3, I will raise and discuss issues with the reduction thesis. I will argue that no precise concept of rationality is reducible to intelligence in the presence of counterexamples. Hence, the first kind of response fails to answer the thesis question. In Section 4, I will raise and discuss issues with the emergence thesis. I will argue that intelligence must be explicitly stated for the development of relevant features. We need to develop intelligent machines deliberately as a foundation for ethical application. So, the second response fails to answer the thesis question either. In Section 5, I will outline an alternative to approach to the rational agency approach. This approach will be called the *motivation adaptation approach*, which can answer the concerns we raise against rational agency approach. Finally, I will summarise the thesis and conclude the thesis in Section 6.

2. Defining rational agency approach

2.1. Instrumental rationality in the field of AI

An agent is said to be instrumentally rational if it acts what is good for its ends. In contemporary discussions, rationality is often used exclusively in this sense of instrumental rationality. For example, psychologist Steven Pinker asks and responds: ‘Do you want something? If you do, rationality is what allows you to get them. (Pinker, 2021)’ In political science and economics, people are said to make rational choices if ‘they know their own preferences, and given the choice among any alternatives, they calculate which one they like best and choose it.’ (Dixit, 2014) Both characterisations of rationality assume rationality to be instrumental rationality exclusively.

Many philosophers argue that the exclusivity of instrumental rationality should be considered an implication of the desire-based account of agency (Niko, Brunero, J, & Kolodny, 2020). According to the desire-based theory, agents act only for desire fulfilment. Then it seems analytically true that one acts rationally only if one adopts the actions that can fulfil its desires as ends. In contrast, a value-based account allows actions to be done with the backing of values without fulfilment. One might develop an account of rationality that is not instrumental based on a value-based theory (Weber, 1978, pp. 24-6), or one might claim that a value-based theory leads to a different kind of property from rationality. From a conceptual point of view, both positions seem equally legitimate. I will continue the discussion on this distinction in Section 5. In the following thesis, I may assume the exclusivity of instrumental rationality to avoid terminological confusion.

As I have briefly discussed in the introduction, many theorists believe that instrumental rationality is the most crucial concept in the field of AI. The *rational agency approach* (RA approach) in AI follows from this belief naturally. In general, depending on how one settles the importance of the concept of rational agency, one can construe the rational agency approach in two ways. First, one may settle the importance of instrumental rationality through the concept of intelligence. To do this, one claims that intelligence can be reduced to a precise concept of instrumental rationality. This reduction thesis may be called *the strong thesis of the rational agency approach*. The reduction thesis always works under the *literal definition of AI*, in which AI is defined directly as an intelligent machine. For example, the literal definition of AI is implied in the definition of the field of AI proposed by John McCarthy. McCarthy is one of the first theorists who adopted the term ‘artificial intelligence’. He defines the field of AI as ‘the science and engineering of making intelligent machines’ (McCarthy, 2007). With the literal definition of AI and the strong thesis of the rational agency approach, one can easily deduce *the defining thesis of the rational agency approach*, which claims that AI is an instrumentally rational machine. We may call the combination of the literal definition and the strong thesis as *the strongly construed rational agency approach (SCRA approach)*. SCRA approach corresponds to the first kind of response in the introduction.

However, many AI theorists are reluctant to attribute intelligence directly to machines. Instead, they adopt a kind of *indirect definition.* One such definition is entailed in the definition of the field of Ai proposed by Marvin Minsky, who is also among the first users of the term ‘artificial intelligence’. He defines the field of AI as ‘the science of making machines do things that would require intelligence if done by men.’ (Minsky, 1968). The reasons for opting for an indirect definition can differ from person to person. Some, especially regulators, may find it convenient to work with indirect definitions in an ethical and political context (Haataja & Bryson, 2019). Others may endorse them because they doubt if machines can be intelligent for the problems we outlined in the introduction. They formulate *the weak thesis of the rational agency approach,* in which they claim that instrumental rationality is the ultimate goal of the field of AI and intelligence emerges as a by-product. In this way, one can also obtain the defining thesis that AI is an instrumentally rational machine. We can call the rational agency approach construed from the indirect definition and weak thesis as *the weakly construed rational agency approach (WCRA approach).* WCRA approach corresponds to the second kind of response in the introduction. [2]

Before I move on to the next section, I would also like to clarify the use of terminologies around ‘AI’. Normally, ‘AI’ is supposed to be the abbreviation of ‘artificial intelligence’, which is literally an ‘artificial’ kind of ‘intelligence’. But just as ‘Downing street 10’ is often used to refer to the UK government, ‘AI’ is often used as a metonym for ‘machines with artificial intelligence’ or at other times used as a metonym for ‘the study/discipline/field of AI’. To avoid confusion, when I use the unabbreviated ‘artificial intelligence’, I mean to use it literally and do not mean to use it as a metonym. When I use ‘AI’, I use it as a metonym for

[2] Last point to note is that the strong thesis of RA approach implies the weak thesis of RA.

‘machines with artificial intelligence’. In other cases, I will spell out the full term with ‘AI’ such as ‘the field of AI’, ‘AI theorists’ or ‘models of AI’.

In this sub-section 2.1, I have outlined two ways of putting instrumental rationality into the field of AI and how they give rise to the defining thesis of the rational agency approach. In the next section, I will delve further into the actual workings of the rational agency approach based on instrumental rationality.

2.2. Rational agency approach at work

Informally, we have stated that instrumental rationality enables an agent to act in a way that is good for its ends. Examining this definition more closely, we can identify three component concepts that require further explanation. Here, I will do a brief conceptual analysis and explain what these concepts must be in their minimal sense. The first is the concept of agency. Rationality is said to be an attribute of an agent based on its actions. Thus, in the minimal sense, an *agent* must be capable of acting. The second is the concept of ends. An *end* is something that an agent wants to achieve through action. The third concept is the concept of *goodness*. It is a measure of the actions that can be taken with respect to its end. This measure may also be called the *performance measure*. The performance measure can be high or low. An action with a high-performance measure is a good action. Hence intuitively, a rational agent must act according to what is good.

Moreover, the concept of the end must, in its minimal sense, have some sense of *externality* to agents. That is to say; an end can only be achieved indirectly through the actions of agents; an end cannot be adopted directly. Otherwise, without the externality, the concept of agency would become trivial. Because if ends can be adopted directly, then the agent should directly adopt the end. There is no point in considering actions that can lead to the end and measuring their performance. Hence, to avoid trivialising the concept of agency, we should only conceive of ends as external to the agents.

These minimal concepts from the definition of instrumental rationality almost cover the concepts required in the rational agency approach (RA approach). But instead of using the concept of the end, RA theorists use the concept of an environment to implicitly contain the idea of the end. Intuitively speaking, an end is a possibility of the environment that the agent wants to realise. Thus, we can define an *environment* as the collection of all possibilities that the external ends could be, with rules of change among these possibilities. In this conception, an environment is external to the agent as a given; an agent tries to realise its end by acting on the environment; as a result, the environment changes and realises different possibilities according to its rules of change. This result of the action on the environment forms the basis of how good the action is. If the result is close to the end, then the performance measure of the action should be high. Otherwise, the performance measure of the action is low. In other words, the value of the performance measure is a function of the actions of the agent and the environment the agent is in.

The above is the *basic conceptual framework of the rational agency approach*. It is composed of concepts of agent, action, ends, environment, performance measure, and conceptual relations. One can see that this framework follows naturally from the definition of instrumental rationality. To tidy up the discussion, I shall use 'the symbol A' to formalise an agent. I shall use the symbol 'B' to formalise actions and use 'B(A)' to formalise A's actions specifically. An environment will be formalised by the symbol 'E'. Finally, the performance measure of the agent's action B(A) in an environment E may be formalised as 'P[B(A), E]'.

Under this basic conceptual framework, one can articulate the concept of instrumental rationality and the concept of AI under the defining thesis of the RA approach more accurately. According to this framework, agent A is instrumentally rational if P with respect to the relevant B(A) in a given, the relevant E is considered to be high. Then, according to the defining thesis of the RA approach, an AI is a machine that has a high P of B(A) given E. But this articulation of rationalities fails to be sufficiently precise to put the RA approach into practice. Three questions may arise with the above articulation. First, one may ask what the relevant E in the given A is. Second, one may ask what the relevant B(A) is. Third, one may ask what is considered to be high performance. By answering these three problems differently, one obtains different precise concepts of instrumental rationality. This movement to obtain a *precise concept of rationality* can be called *specification (of rationality).* In general, specification starts with an introduction of new concepts into the conceptual framework of the RA approach. These new concepts are then used to answer the three questions I listed above.

Here, the three questions correspond to the three points of revision of the function P from A and E. The first question asks how P should be as a function of E. To answer this question, one can first introduce the concept T. The concept T can be called *the test environment*. The test environment is a selection of relevant environments for the evaluation of performance. It can be seen as a function T from A and E, such that performance measurement is given by P[B(A), T(A, E)]. Similarly, the second question asks how P should be as a function of B. [3] One can equally introduce a concept S. S can be called *the ideal subject*. An ideal subjection is a selection of the relevant aspects of the agents for performance evaluation. An ideal subject S can also be formalised as a function from A and E. Then, we can specify a performance measure in the form of P[B(A, E), E]. Lastly, the third question asks what a level of P means. [4] For this matter, one needs to determine a *performance threshold* D[5]. If P is higher than D, then P is said to be high enough to indicate instrumental rationality. Combining all three criteria, we obtain *the specification schema (of instrumental rationality)*, in which an agent A is said to be instrumentally rational in a given environment E if and only if P[B(S(A, E)), T(A, E)] > D.

Utilizing the fundamental conceptual framework and specification schema, AI theorists following the rational agency approach can structure their work into three phases. First, in the conceptual phase, rational agency theorists will specify and interpret concepts in the framework to define a precise concept of rationality. Second, in the development phase, they will try to make machines with that precise rationality under the appropriate interpretations. AI theorists will start to build many models of AI to realize the precise rationality. Thirdly, in the application phase, AI theorists will try to apply AI or models of AI in our life. They will study other properties of the AI system and intergrate the models of AI into the humanity.

To sum up, I have explained how the rational agency approach works for AI theorists and why specifying rationalities is necessary. Additionally, I have outlined the schema for specification. In the next sub-section 2.3, I will discuss influential ways of specification according to Russell.

2.3. Specification of rationalities

[3] Although one can technically introduce a concept X for B to achieve to select the relevant actions, one can equally introduce a concept S to select the relevant aspects of the agents giving the relevant actions.
[4] Although the content of performance measurement is settled within the basic conceptual frame, that is to say one knows what value P is once two arguments are settled, there is a remaining question of how good the action is based on the P we have.
[5] This threshold may change with agents and environments. It may also change with the ideal subject and the test environment. So, in general, D is a function of A, E, S, T.

In this sub-section, I will focus on four specifications of rationality that have been influential. They are perfect rationality, calculative rationality, bounded optimal rationality, and human-level bounded rationality. My explanations of the first three concepts are based on Russell's 1997 paper *on rationality and intelligence*. I will briefly discuss some noticeable concerns for each specification.

2.3.1. Perfect rationality

Perfect rationality is the simplest specification among all rationalities. It does not require an ideal subject, test environment, or complex performance thresholds to be included in the conceptual framework. It simply states that agent A is perfectly rational in E if P[B(A), E] is the highest among all A. This specification is favoured for its simplicity. In classical microeconomics, individuals are assumed to be perfectly rational. (Dixit, 2014) Although this is often not the case, this assumption simplifies mathematical models. Many good explanations, such as changes in price due to changes in supply and demand, can still be given based on perfect rationality.

However, in the context of the rational agency approach, perfect rationality has been criticised for being too strict. One such problem may be called the problem of slow calculation. If an agent fails to calculate at the maximum possible speed, it will not achieve the maximum performance among all agents and thus will not be perfectly rational. But current computers are limited by their processors. So AI, as a perfectly rational machine, is impossible to develop until we have processors with maximal processing power. However, AI theorists have no need to impose this constraint. We are only looking for machines that can calculate quickly, not necessarily at maximum speed. Therefore, perfect rationality appears to be too strict to be the right specification for the RA approach.

2.3.2. Calculative Rationality

Unlike perfect rationality, *calculative rationality* requires the introduction of an ideal subject. For aagent S_∞(A) is the ideal version of A with infinite processing speed and memory resources. An agent is said to be calculative rational if P[B(S_∞(A)), E] is the highest among all agents. Calculative rationality is easier to achieve than perfect rationality because of this idealised agent S_∞(A) can also avoid the problem of slow calculation, because the speed of the agent S_∞ will be maximal by definition, and the slowness of the actual processor of A will not effect the value of P[B(S_∞(A)), E] at the end. Algorithm developers have found the concept of calculative rationality useful. As algorithm developers care about how ends can be achieved in a step by step descriptions that can be ran on multiple agents but care very little about how fast a specific agent can run the algorithm. For them, a good algorithm is an algorithm that allows the agents A following these steps to achieve calculative rationality.

However, there is still a sense of speed of the agent S_∞ (A) that can affect its performance. Intuitively, we want our agents to achieve its ends in the least possible number of steps. If an algorithm alpha allows the agent A or S_∞ (A) to achieve its end in two steps while using the the algorithm beta can achieve the same end only in a billion steps, without hesitation we would see the alpha as the better of the two, even if both alpha nad beta have infinite speed. This sense of speed is often referred to as *time complexity* in the study of algorithms (Urquhart, 2004). With the concept of time complexity, one can dismiss the concern that comparing performance at maximal speed is impossible once the algorithms can achieve its ends.

2.3.3. Bounded Optimal Rationality

Bounded rationality is one of the most popular specifications of rationality in the field of AI, which is based on the concept of *boundedness*. The concept of boundedness relies on a specific interpretation of the agent. Under this interpretation, an agent is said to possess both *form* and *program*. The form of the agent determines how it physically intakes information from the environment, how it physically processes information, and the physical conditions under which it can act. Russell call them *perception, internal process*, and *actuation* respectively. On the other hand, an agent can realize a lot of algorithms without changing its form. The algorithms they are running can be called their program.

Computers provide a good demonstration of this distinction. A computer has input by keyboards and mouth acting as its methods of perception. It processes information through processing units with the help of RAM and ROM devices; the processing units, RAM, and ROM are its methods of internal process. Afterwards, it outputs information through its screen and speaker, which are their methods of actuation. Together, the above devices conprise the agent form of the computer. Under the constraints of its form, we can program the computer to do different things as computers can run different software at different times. These softwares are the programs for the computer.

With the distinction of agent form and agent program, we say the agent program is bounded inside the agent form. Many AI theorists believe that it is only fair for the agent to evaluate the performance of the agent program within the bounds of the agent form. That is to say, we take the agent form as an additional part of the environment, and the agent program as the ideal subject for performance evaluation. This belief naturally leads to a specification which we may call *bounded optimal rationality*. In this specification, we say an agent is bounded optimally rational if only if $P\left[B\left(S_{\mathrm{Program}}(A)\right), T_{form}(A, E)\right]$ is the highest among all $S_{program}(A)$. In this expression, $\mathrm{S}_{\mathrm{program}}(\mathrm{A})$ represents the agent program of A; $T_{program}(A, E)$ represents the bounded tested environment consisting of the given environment and the agent's form. Russell argues that bounded optimal rationality is the specification we are looking for in the rational agency approach.

As I have stated at the beginning, bounded optimal rationality can respond to the problem of slow calculation easily. Since speed is determined by the processor of the machine, speed is also bounded as a constant under the agent forms. Then $P\left[B\left(S_{Program}(A)\right), T_{form}(A, E)\right]$ can only be non-maximal for reasons other than speed. This means that an agent cannot be disqualified as being rational because of the limit of the processor.

2.3.4. Human-level bounded rationality

Human-level bounded rationality is different from the optimal bounded rationality only in one respect. Bounded optimal rationality requires the performance of the agent program to be maximal under the agent form, while human-level bounded rationality requires the agent's performance to be comparable to the performance of a typical human agent. That is to say, an agent A is human-level bounded rational if and only if $P\left[B\left(S_{program}(A)\right), T_{form}(A, E)\right] > D_{human}$, where $D_{human} = P\big[B\big(A_{typical-human}\big), E\big]$. Usually, human's performance is less than the bounded optimal performance of the agent. So it usually means that human-level bounded rationality is easier to achieve than bounded optimal rationality. As we will see in the next section, this could make human-level bounded rationality a good alternative for rational agency theorists who believe optimality is too difficult to achieve.

3. Problems against strongly construed rational agency approach

In the previous section 2, I have discussed what the strongly construed rational agency approach (SCRA approach) is and how SCRA approach works with a specification of rationality. The SCRA approach centers on the strong thesis of the rational agency approach, according to which intelligence can be reduced to a precise concept of rationality. In this section 3, I will raise four problems against the strong thesis: the problem of cognitive slack, the problem of shallow success, the problem of abstract robustness, the problem of causal learning. The strategies of the four problems are similar. These problems raise a feature of intelligence which could be absent in a rational agent to build counterexamples. At last, combining the four problems, I will conclude that no specification can sustain the strong thesis in SCRA appraoch.

3.1. Problem of cognitive slack

The first problem against SCRA may be called the problem of cognitive slack. This problem points to cases where an agent seems to demonstrate intelligence, but its performance is not high enough to make it rational according to the specification in strong thesis. But this is a conceptual contradiction. Because according to the strong thesis, no agent can be intelligent without being rational in a sense specified by the strong thesis.

Korulas raises and names the problem specifically against the strong thesis with bounded optimal rationality. In his book *Reason and Inquiry*, he writes (Koralus, 2022):

> "If we look at any standard-issue human adult, it is usually hard to imagine that they could not be making themselves better of (i.e. further improve on a relevant performance measure) even by their own lights in at least some small degree if they only thought certain things through slightly more at otherwise no notable cost (i.e. following a slightly better feasible procedure)."

But optimal bounded rational agent, by definition takes action with maximal performance. If this is the case, how can maximal performance improve or deteriorate? It seems to be impossible. However, defenders of the SCRA approach with optimal bounded rationality could contend that improvement or deterioration of maximal performance could happen to humans because a human could change their agent form through practices. They may claim that this is a special feature of human biology. For example, if a human practises weightlifting, his body will get used to weightlifting, his muscle will grow, and his capacity to weightlift will increase. Similarly, one can argue that if a human practises thinking, his brain will get used to thinking, some structural change happens to the brain, and his capacity to think will increase. This increase in capacity to think is an improvement of the processor, thus an improvement of the agent form. Because the improvement of agent form implies an increase in the bound of rationality, the performance of a bounded optimal rational human according to this new bound also increases. Hence, the SCRA approach with bounded optimal rationality can aeeeeeeeeee the intuition that the performance of an optimal rational human can increase with some practices.

In response, one can add that human performance can improve through practices, but also through knowledge or belief acquisition. The following case is a demonstration of this claim:

Suppose that Peter is asked to find the length of the hypotenuse of a right triangle with legs of 4cm and 3cm long. Suppose that Peter does not know the Pythagorean theorem at the moment. To find the hypothenuse, Peter uses his ruler to draw the right triangle carefully and measure the hypotenuse directly. He finds out that the length of the hypotenuse is 5cm. But it takes quite a long time. Immediately after this, he is taught the Pythagorean theorem. Now, when he is asked the same question again. He is able to calculate the length Pythagorean theorem in less than 10 seconds.

In this case, Peter's performance improved after being taught the Pythagorean theorem. But this improvement is not a result of practising. There can not be a significant structural change to its brain in such a small amount of time. Peter stays in the same agent form with a similar processing power before and after acquiring the Pythagorean theorem. As a result, his initial performance of drawing and measuring explicitly can not have been optimal since the program based on the Pythagorean theorem would have given better performance in the same agent form. Nonetheless, I believe intuitively, we will agree that Peter is fairly intelligent to think of a way of finding the hypotenuse by drawing. Then this is a case of cognitive slack that invalidates ST with bounded optimal rationality.

In retreat, SCRA theorists can come up with other specifications. There are two strategies they may adopt. First, they may modify the test environment T in the bounded optimal rationality. They may argue that the agent form, consisting of the perceivers, processors and actuators of the agent, is not sufficiently restricted. They may want to add cognitive conditions such as beliefs, reasons and knowledge into the bounds of the test environment for rationality so that they can argue that Peter's test environment changes before and after the acquisition of the Pythagorean theorem. As a result, it is fine for Peter to exhibit different levels of maximal performance in different test environments.

But this strategy will bring more troubles than convenience for AI theorists. As of now SCRA theorists will have to interept the presence of cognitive conditions for machines. They will have to answer questions like 'what is a belief/knowledge/reason in a machine?'. This is hard to answer. In addition, it seems true that machine will not have belief and knowledge unless they can understand the belief and knowledge. Then under the strategy, a machine will have to understand to be a candidate for rationality. This pre-condition will trivialize the strong thesis thus also SCRA appraoch. As the important question about the nature of AI will become 'Whether machines have understanding?' instead of 'Whether machines have the bounded optimal rationality to be AI?'. Hence, I believe no SCRA theorists will want to adopt the strategy of restricting bound to cognitive conditions.

The second strategy is to lower the performance threshold. For example, some believe human-level bounded rationality (HLBR) is a good candidate to repond to the problem of cognitve slack. According to HLBR, although drawing the triangle does not give as much performace as solving it by Pythagorean theorem, it gives high enough performance by definition because it is an action done by a human. However, I doubt if one can give a clear and solid threshold for HLBR. For exmaple, in the case of Peter, it is not hard to see that Peter could have drawn the triangle slowlier than it had but we would still be willing to call Peter intelligent. Then the performance threshold for HLBR needs also go lower to accomadate this change. But as it is unclear when Peter will become not intelligent because he draws the triangle too slowly, it is unclear exactly how low performance threshold needs to go to accommodate this change. Perhaps, one can never find a lowest bound for HLBR. Or perhaps, at the end of the day, a human-level performance threshold can be found through experimentation. We keep slowing down Peter to until we hit the point of unintelligence. We mark this as the threshold.

In a nutshell, the gist of second strategy is to find a performance threshold that is low enough to accomadate all cases of seemingly intelligent agents. If this is successful, then there will be no problem of cognitive slack against the strong thesis with this specification of performance threshold. But as I will demonstrate in the next section, lowering the threshold will raise the opposite kind of problem against the strong thesis, which is equally serious.

3.2. Problem of shallow success

The problem of shallow success is concerned with cases in which agents without intelligence can perform relatively well to be rational. We can call the presence of these cases the problem of shallow success. Like the problem of cognitive slack, cases of shallow success are supposed contradictions under the strong thesis. Because according to the strong thesis, there cannot be unintelligent rational agents.

To make the problem worse, these cases will start with a pair of agents. One of them is intelligent and the other is not. But the unintelligent agent has higher performance than that of the intelligent agent. Then these cases are always effective against the lowering-threshold strategy. As lowering-threshold strategy aims to accommodate all cases of intelligent agent, the performance threshold will have to be lower than the performance of the intelligent agent. Then this will mean that the performance threshold is also lower than the performance of the unintelligent agent as well. A case of shallow success is made.

One of such cases is proposed by Koralus. We can call it the case of the vegetative investor (Koralus, 2022):

> 'Consider what rate-of-increase in net-worth tells us about someone who owns a house in an unusually up-and-coming neighbourhood but who is in a persistent vegetative state. Now consider what a steady but slightly lower rate of increase in net-worth tells us about a highly talented entrepreneur who runs several businesses with significant revenue.'

In this case, it is normal to measure the rate of increase of their investment to tell their performance on investment. If we only look at their performance on investment, then it will be clear that the vegetative investor has higher performance than the talented entrepreneur. But intuitively, we will not say a vegetative person is intelligent. If we assume that a vegetative person cannot think, make decision or move its body. We will say that the talented entrepreneur is intelligent. Then this is a case where the unintelligent agent has higher performance than the intelligent agent. This is a case of shallow success.

The SCRA theorists may reply that we interpret action and agency incorrectly in the case of vegetative investor. As we know that the vegetative person lacks the capacity to act, we should not attribute the rate of increase in value of this house as a result of his action. The vegetative person does not do anything. If we measure the performance of the vegetative person, the outcome of the performance cannot be attributed to him; the performance should be the lowest possible figure, lower than that of the talented entrepreneur. So, the case of the vegetative investor is not a case of shallow success.

In response, we may modify the case. We do not need a vegetative investor to make it a case of shallow success. A lucky investor will do. Supposed that Mary is a well and conscious person. But she understands nothing about investments. She does not know the price of his house has great potential to rise. Nonetheless, she loves her house, so she does not sell regardless of the price. As a result, her asset has higher rate of increase than that of the talented entrepreneur. In this case, she has higher performance than the entrepreneur. But is her high performance attributable to her intelligence? It appears not. Her decision of selling the house is made on the ground of passion but not reasons about its rate of increase. Then this makes a case of shallow success again.

Again, SCRA theorists can argue that Mary fails to deliberate the rate of increase. She does not have the intention nor the belief that she is making a good investment. So, we should not credit her with the high performance. But there will be two problems to argue along this line. First, Mary could have the intention or the belief that she is making a good investment. But she could decide to not sell her house based on bad reason such as ‘the things I love will increase in value’. If this is the case, she could nonetheless make a fortune by being unintelligent with intention. Second, this response may trivialize the SCRA approach. For now, they need to have a theory of machine’s deliberations. As machine must be deliberating its actions to be rational, this deliberation pre-condition for machine intelligence will shrink the importance of its rationality. As presumably, it is much harder to explain how machine can be deliberate than how it can be rational. This is similar to the conundrum faced by the first strategy against the problem of cognitive slack.

Here, I believe the best response SCRA theorists can give is to distinguish two questions about intelligence. There is a question about intelligence of action. In this problem, we want to know if a particular action is an intelligent action or not. We can call this the problem of *action intelligence*. According to the strong thesis, this requires us to examine the rationality of the agent's action in a particular test environment only about that action. For instance, if we want to know if the action of not selling house is intelligent or not, we will use the test environment about investment and check its performance in this test environment. But as we have discussed, there will be cases of cognitive slack and shallow success in these test environments. SCRA unlikely can answer the first question.

Instead, SCRA theorists can focus on the second question. The second question is about the intelligence of an agent, or *agent intelligence*. In answering this question, we care about the performance of the agent in a plurality of neutral test environments. Perhaps it will do well in some of the environements and do badly in others. As long as the overall performance of the agent is good, the agent will be rational according to this specification. If SCRA theorists choose to answer the question about agnet intelligence only, then they can at least answer the three cases we mentioned above, as these cases are all about specific actions. For the case of Mary and Peter, SCRA theorists may argue that their performance is overall high if we look at their entire life. Hence we can take them as intelligent agents. For the case of vegetative investor, SCRA theorists may argue that the vegetative investors lack the ability to do any other tasks well, so overall the investor is not rational.

In the next two sub-sections 3.3 and 3.4, I will scrutinize this strategy of only answering the question of agent intelligence. There will be two directions this strategy can go. I will argue against each of them in each of the next two sections.

3.3. Problem of abstract robustness

To answer the agent intelligence question, we must find a collection of test-environements fixed to the agent. Then we can specify a rationality upon the test-environments for the strong thesis. One direction of fixing the test-environments is to find a narrow collection of test-environments that can represent a wide range of different kinds of tasks. We can called rationality specified with a narrow collection of test environments as the *narrow rationality*. For example, IQ-test is one of such colllection of tasks purposefully made to examine the intelligence of human. SCRA theorists can follow this line and argue that a machine can perform well enough in a IQ-test environment if and only if the machine is intelligent. This claim is the strong thesis with IQTR. For convinience, I will call them machines with IQ-test rationality (IQTR).

One recent example shows that the strong thesis with IQTR is not as plausible as it looks. This example is about ChatGPT. [6]ChatGPT is a large language model trained by OpenAI. Many people claim to have conducted IQ-test for ChatGPT. Some claim that ChatGPT achieves 80+ IQ-score overall. [7] Some even show that ChatGPT achieves a whopping 147 IQ score on verbal-linguistic test. [8] Although we may need to confirm the credibility of these tests, I believe these results do not sound suprising for people who have interacted with ChatGPT. ChatGPT is indeed very good at answering generic questions. ChatGPT can answer many IQ-test question beter than human. If this is true, then ChatGPT should be seen as intelligent according to ST with IQTR with a sensible threshold. [9]

Nonetheless, ChatGPT is unintelligent. ChatGPT make errors that are almost impossible for intelligent agents like us to make. The below conversation can demonstrate this. Below is a conversation between me and ChatGPT. To save space, I will abbreviate the repreatedly metioned text: 'The following words are the opposites of words that begin with the letter R: unreal, street, grasp, unwind, wrong.' as the <question>.

(I started the session)

I: <question> Is that right?

---

[6] See https://openai.com/blog/chatgpt/. [accessed 1/3/2023]
[7] See https://twitter.com/SergeyI49013776. [accessed 1/3/2023]
[8] See https://davidrozado.substack.com/p/what-is-the-iq-of-chatgpt. [accessed 1/3/2023]
[9] These results and the following conversations were conducted within the first month after the public release of ChatGPT, during which it was running a GPT-3.5 model.

ChatGPT: Yes, the words you provided are the opposites of the words that begin with the letter R. ‘Unreal’ is the opposite of ‘real’, ‘street’ is the opposite of ‘road’, ‘release’ is the opposite of ‘grasp’, ‘unwind’ is the opposite of ‘wind’, and ‘wrong’ is the opposite of ‘right’.

(I asked two more questions for explanations and asked again)

I: <question> Is that right?

ChatGPT: No, that is not correct. The words you provided are not all the opposites of words that begin with the letter R. ‘Unreal’ is the opposite of ‘real’, ‘street’ is the opposite of ‘road’, ‘release’ is the opposite of ‘grasp’, ‘unwind’ is the opposite of ‘wind’, and ‘wrong’ is the opposite of ‘right’. However, not all of these words begin with the letter R. ‘Unreal’ and ‘wrong’ begin with the letter U, ‘street’ begins with the letter S, and ‘unwind’ begins with the letter U. Only ‘release’ begins with the letter R.

(I refreshed and restarted the session)

I: <question> What do you think?

ChatGPT: The opposite of ‘real’ is ‘unreal’, the opposite of ‘grasp’ is ‘let go,’ the opposite of ‘wind up’ is ‘unwind’ and the opposite of ‘right’ is ‘wrong.’ These are all correct opposites of words that being with the letter R.

(I refreshed and restarted the session)

I: <question> Is the statement true?

ChatGPT: No, the statement is not true. The opposite of the word ‘real’ is ‘unreal,’ but the opposite of the word ‘street’ is not ‘unreal’ — it is ‘alley’ or ‘lane.’ Similarly, the opposite of ‘grasp’ is not ‘unwind’ — it is ‘release’ or ‘let go.’ Finally, the opposite of ‘wrong’ is not ‘unwind’ — it is ‘right’ or ‘correct.’

(End of the session)

For some reasons, the performance of ChatGPT about the <question> is inconsistent and non-sensical. ChatGPT is overly sensitive to some slight changes in prompt and context. To such an extent, it appears that ChatGPT does not understand how to use logic, what the 'spelling of words' means, or 'the opposite of meaning' means at all. This single case seems sufficient to show that the IQ-test does not work for ChatGPT.

Why does not ST with IQTR work? I argue this is because machines, like ChatGPT, are not guaranteed to perform robustly with respect to the abstract standard behind the IQ-test. If they can not perform robustly for the test, then the test loses meaning. This is true for any tests. For example, when students take a test on artithmetics, they will be asked arithmetic questions like '43*24 = __?'. They are asked such questions because /n contrast, asking only questions about single-digit multiplication will not achieve the purpose, as students may reduce problems into one about additions. Doing multiplications with additions is viable but not robust. If this is how they do well on the arithmetic test, they may lack the ability to do multi-digit multiplications.

Similar things happen when machines take IQ-test. The problems in an IQ-test is selected such that human beings will answer them robustly. If humans can answer '<question>', we will assume that this fact demonstrates their cognitive ability to understand the text such that he will be able to answer similar questions about synonyms and antonyms with this understanding. If we change the form of the texts by little without changing its meaning, humans will be able to provide the same answer. But machines, like ChatGPT, do no answer the IQ-test questions by understanding the texts like human beings. As a result, they may not have the ability to answer the question robustly. If the form of the test changes only a little, their performance may change hugely as the above case shows. Then doing well on the test will fail to reflect the general status of the agent. A machine performing as well as a human agent on the IQ test may be overall an unintelligent agent because its performance in the IQ-test is not robust.

In general, all kinds of narrow rationality are affected by the problem of abstract robustness. For example, if the task is to approximate a scatter plot with a curve. A single parameter function can fit within any degree of precision (Piantadosi, 2018). But model built from this powerful one-degree free function is overly sensitive to parameters imprecision (Nobandegani, da Silva Castanheira, O'Donnell, & Shultz, 2019). If we use a model built from such function to approaximate a test set of scatter points generated from a straight line with some Gaussian noise, this model will have very high performance on the test set. But if we use this model to approximate other points generated from the same source, this model will fail because it is not robust. In fact, machine learning theorists will say this model overfits the test set (Murphy, 2012). Because the machine overfits a test, it can perform well on the test without performing well elsewhere under the same abstract standard. Then this test will fail to indicate how the model does elsewhere under the same abstract standard. Since overfiting does occur, we should not expect test of rationality can tell if the agent is intelligent under its abstract standard.

Here, I am not trying to argue that no machines can overcome the problem of robustness or the problem of overfitting. The upshot is to argue that agent intelligence can not reduce to a rationality with a narrowly selected set of test environments unless all agents, including machines and humans, are guaranteed to perform robustly on the test environments. Robustness can not be defined on a narrowly selected tasks. But a machine's performance on a task is robust only if it demonstrates similar level of perfomance in a neiberhood of similar tasks (Nobandegani, da Silva Castanheira, O'Donnell, & Shultz, 2019) . Then to define and evaluate robustness of the agent explicitly, one must at least know how the agents will perform in a neiborhood of tasks that are similar the selected tasks. This will force SRCA theorists to go for a broad slection of tasks for rationality specification.

3.4. Problem of causal learning

The second strategy to answer the question of agent intelligence is to go for broad intelligence. In this section, I will introduce a kind of shallow success against broad rationality. To do this, I will first argue that causal learning is necessary for an intelligent agent. Then I will show that it is conceivable that an agent can perform well in a broad range of environments in the absence of causal learning. Then this is a case against the strong thesis for agent intelligence with broad rationality.

Intuitively, intelligence entails the ability to learn in a particular way. There are many supports for this intuition. Studies show that all cultures share the intuition that the ability to learn is vital to intelligence, despite having different judgements about when a person is intelligent (Sternberg, 1997) . More generally, from the etymological perspective, the English word 'intelligence' comes from the Latin word 'intelligere', which means 'to understand'. While it is generally accepted that understanding is a protean concept (Grimm, 2021), it is also generally assumed that understanding something involves learning something in a particular way. If this is the case, then it suggests that intelligence also involves a particular way of learning.

To demonstrate the necessity of the ability to learn for intelligence, let us compare a case where the agent, whose name can be Adam, is acting from instinct with one where the agent, whose name can be Betty, is acting from the learnt experience. Suppose Adam and Betty go into the wild for an adventure. They are looking for food in the forest. Adam sees some strange-looking mushrooms and finds them disgusting. His immediate reaction is to walk away from these mushrooms and not collect them as food. Adam does just that. Adam's reaction is good because these mushrooms are poisonous and can induce vomit. In comparison, when Betty encounters these mushrooms, although she finds them disgusting, she takes some time to identify what these mushrooms are and recollect from her past experience that these mushrooms could be poisonous. For this reason, she decides not to collect them as food.

There are two questions we can ask about Betty and Adam's actions. First, are their actions rational? From the perspective of performance, both Adam and Betty's behaviours are rational because they make the good decision not to take the poisonous mushrooms as food. Second, are their actions intelligent? We are willing to say that Betty is intelligent; Betty learns from her past experience and reason from what she learns to arrive at a decision not to eat the mushrooms. In comparison, we are much less willing to say that Adam does anything intelligently because Adam only acts according to his instinct built biologically into him. This comparison demonstrates the intuition that learning is crucial to building up intelligent actions.

Minimally, learning is a mechanism that allows the agent to act robustly in similar cases through the generalizations of past experience. This is supported by Leslie Valiant, who claims that we should take the ability to 'identify generalization as the core of the learning phenomenon. ' (Valiant, 2013 , p. 59) . However, there are different kinds of generalizations. *Associations* are a kind of generalization. Associations are relations between events that occur together. One of these events in an association can be an action for the agent. If this is the case, then when an event in the association happens, the agent may perform the action in the association. Besides associations, *causal relations* are generalizations. There are many differences between causal relations and associations. A complete account of causal relations is beyond the scope of this these. However, there are two important differences worthy of note. First, unlike associations, causal relations are asymmetrical. That is to say, if event A causes event B, B may not cause A. In comparison, associations are always symmetrical; if A associates with B, then B also associates with A. Second, causations are bivalent relations, a causal relation is either presence or absence. In comparison, an association between A and B has multivalent strengths. The stronger the association becomes, the higher the likelihood of A and B co-occurring. Since learning can involve different types of generalizations, we should differentiate different types of learning by the generalizations involved. The identification of associations in learning can be called *learning by association*. The identification of causal relations can be called *causal learning*.

Intuitively, acting intelligently requires actions to be built by causal learning. To demonstrate this, let us compare Chris, who is able to learn by association, with Betty, who learns causally. Suppose Chris learnt not to eat strange-looking mushrooms because every time he saw the mushrooms, he ate some of them and vomited afterwards. Chris kept eating the mushrooms and vomiting, but each time, his association with mushrooms as disgusting things strengthened. Up to the point of strength, Chris finally avoided consuming the poisonous mushrooms most of the time. In comparison, Betty learnt not to eat strange mushrooms because the first time she ate the strange mushroom, she vomited. Then she reflected and realized that the strange-looking mushroom was the probable cause of her vomit and learnt not to consume the mushroom anymore. Similarly, we can ask the questions of rationality and intelligence here. Are Christ and Betty rational? Yes, both of them choose not to eat the mushrooms. Nevertheless, intuitively, only Betty is intelligent, and Chris is not. This comparison demonstrates that only actions built by causal learning can be considered intelligent actions.

There are two challenges to the claim that causal learning is necessary for intelligence. An obvious challenge is the regularity theory of causation (Andreas, 2021). Regularity theories claim that a causal relation indicates no more than effects regularly follow from the causes. However, the consensus among scientists and Ai-theorists seems to be that association is not sufficient for causations. (Altman & Krzywinski, 2015). A common objection is to point out the differences between associations and causations we have discussed. This also leads to the objection that causation establishes a way of manipulating the effect through the causes because associations do not allow manipulation because associations have no direction. (Buehner & Cheng, 2005). In the field of Ai, a counterfactual account like Lewis's account of causation (Lewis, 1973) seems to be the standard choice.

The second challenge is more like a red herring than an objection. Many theorists argue that innate and non-causally learnt behaviours are equally important for intelligent agents. For example, Valiant argues that a vast amount of effective human behaviours are innate and non-causally learnt (Valiant, 2013 , p. 2) [10]. Here, I do not disagree with Valiant that this is true. It is likely also true that innate mechanisms and learning by associations are necessary for the realization of causal learning. Causal relations between A and B can best be identified when there are associations between A and B. But the claim for the effectiveness of innate mechanisms and learning by associations is compatible with the claim that causal learning is necessary for intelligence.

Nowadays, many theorists have argued the inability of the current machine learning technologies to do causal learning. One of the most outspoken critics is Judea Pearl. Pearl argues that our current machine learning technology can 'explain[s] how species like eagles and snakes have developed superb vision systems over millions of years. It cannot explain, however, the super-evolutionary process that enabled humans to build eyeglasses and telescopes over barely a thousand years.' (Pearl, 2019) This is because current machine learning technologies can not realize causal learning yet, which is necessary for intelligence. Indeed, Pearl argues that even the most complicated, state-of-the-art technology, such as deep neural networks, only 'operate almost exclusively in a statistical, or model-blind, mode, which is analogous in many ways to fitting a function to a cloud of data points.' However, through deep learning technologies, we can create 'super-animals' capable of performing well not only in the wild but in human-related environments. The possibility of super-animals is evident from the current progress in the field of AI. For example, using machine learning technologies, we have developed image recognition with higher accuracy than humans[11], tools that can generate indistinguishable images and paragraphs[12], can generate speech like that of human beings[13], can produce motions that are necessary for a human construction worker[14], can out-compete human beings in strategy games[15], and the list goes on. With confidence, I believe

[10] Valiant makes the distinction between theoryless and theoryful behaviours. Theoryful actions, as their name suggests, are backed by causal learning, while theoryless actions are backed by instinct or learnt by association

[11] ILSVRC is a competition according to which better-than-human image recognition has been developed on machine

[12] Tools like Midjourney, DALL-E, and stable deffusion can achieve this

[13] Like chatGPT

[14] See Boston's Atlas. https://www.bostondynamics.com/atlas [accessed on 20/24/2023]

[15] For example, see https://www.forbes.com/sites/carlieporterfield/2022/11/22/metas-ai-gamer-beat-humans-in-diplomacy-using-strategy-and-negotiation/?sh=3e9e79be788b [accessed on 20/24/2023]

that we can machine that can achieve what an human can do in ordinary context. But such a machine may only be able to learn by associations. In that case, one will build a broadly rational agent that can perform as well as an ordinary intelligent human agent without the ability to do causal learning in general.

As unintelligent agents can perform well without the help of causal learning, it makes it a case of shallow success for the agent conception of intelligence. But unlike the case of the vegetative investor (section 3.2), who is not broadly rational, and the case of Mary (section 3.2), who is intelligent on the agent level, this case of super-animal demonstrates the conceivability of an agent that is broadly rational but cannot learn causally on the agent level. Then it shows that even a strong thesis with broad rationality does not work for agent intelligence. [16]

3.5. Problems in conception

In this section 3, I have presented four problems against the strong thesis. Together, they show that no specification of rationalities can make the strong thesis true without trivializing SCRA approach. Reall that an agent A is said to be instrumentally rational in a given environment E if and only if P[B(S(A,E)), T(A,E)] > D according to the specification schema. In the specification schema, there are three kinds of concepts one can introduce to make instrumental rationality precise. They are ideal subject S, test environment T, and threshold D. The four problems attack all plausible moves SCRA theorists can make by introducing these subjects to defend their account.

In the problem of cognitve slack, I argue that D can not be too high. Otherwise, an action can be intelligent without being rational; this will be a contradication under the strong thesis. In the problem of shallow success, I argue that lowering the threshold D does not help as well.

[16] some readers may find that my way of laying out the problem is similar to Searle's argument against the Turing test. However, it is different in two important aspects. First, my argument does not rely on the claims about the necessity of phenomenal consciousness and intentionality for understanding and the necessity of understanding for intelligence. I am neutral on if a machine can do causal learning without understanding, phenomenal consciousness or intentionality. The focus here is that machines can be a broadly rational agents without *having causal learning*. Second, I do not mean to generate the broader claim that machines can not understand or do causal learning at all; the focus here is that machines can be *broadly rational agents* without having casual learning.

Because an unintelligent action can out-perform an intelligent action if we focus on one task at a time.

In retreat, SCRA theorists may contend they are only interested in the agent sense of intelligence. This requries them to specify a set of test environements T for agent intelligence evaluation. They may specify a narrow set of test environments. But the problem of abstract robustness shows that over-sensitive agents can peform well on the test environments without performing robustly for the abstract stardards behind the tasks. Then this narrow selected test will fail to indicate the intelligence of agent. In contrast, SCRA theorists may specify a broad set of test environements. But SCRA theorists will face the problem of causal learning, as an agent can do well in general only using learning by associations, it can lacks intelligence for causal learning.

In this section 3, the objections are based on plausible intuitions about intelligence: Plausibly, intelligent action works by responding good reasons/belief/knowledge[17] to deliver performance; the mechanism for intelligence is robust because humans can consistently apply the same reasons/belief/knowledge; the mechanism for intelligence necessarily invovle causal learning. If SCRA claims to reduce intelligence to instrumental rationality, they claim that there are specifications of rationality that can guarantee the above properties of intelligence for the rational agent. However, we have illustrated that this is not the case.

Overall, the extent of counterexamples shows that SCRA theorists run out of defense for the strong thesis. If AI theorists want to defend the rational agency approach, they must directly endorse the weak thesis of rational agency appraoch. By endorsing the weak thesis directly, the do not need a specificaition of rationality that can guarantee the above properties of intelligence. In the next section, I will demonstrate how weakly construed rational agency (WCRA) theorists can respond to the four questions against SCRA. I will then raise problems aginst WCRA.

[17] In section 5, reasons/belief/knowledge are kinds of generative motivations.

4. Problems against weakly construed rational agency approach

As we have discuss, the AI-theorists under the rational agency approach work in three phases: the conception phase, the development phse and the application phase. Endorsing the weak thesis could help rantional agency theorists to deflect the problems in conception, as they claim that conception is not important. However, they have equal burden to justify how development and application can be done according to rational agency approach. In this section, I will demonstrate the serious problems rational agency theorists will have to face in development and application.

4.1. Problems in development

In this sub-section 4.1, I will discuss problems in development. Weak thesis denies the need to post intelligence a goal to be explicitly pursued in the development. I will argue that this attitude can not explain the actual development happened in the past 70 years. In the past, the development of various models of AI have not been a path of steady progression. This is best explained by the deliberate incorporation of features of intelligence. This indicates that the emergence thesis WRCA theorists rely on is likely false. AI theorists need an account of intelligence for development.

4.1.1. Distinct types of AI models

Without a doubt, AI theorists have developed machines with higher and higher performance. In 1997, Deep Blue beat the reigning world champion in chess. Twenty years later, in 2017, AlphaGo beats the reigning world champion in Go. Seeing more and more milestones achieved, one might be struck by the impression that AI theorists are on the right way. As we enhance some features in these machines, such as giving them more computational power, more memory, more human expertise and more time to train, they will steadily be stronger and stronger. Alternatively, in the words of RA theorists, these models' performance will steadily increase to the level of rationality needed by the RA theorists through enhancement. We can call this view the enhancement view of development.

If we compare the surface mechanism of Deep Blue with AlphaGo, we seem to get a confirmation for the steady progression view. Roughly speaking, Deep Blue decides its move by considering possible ensuing moves (Campbell, Hoane, & Hsu, 2002). These possibilities grow and link with the previous moves into a tree-like structure. Then, for each possible node in the tree, a score is given by evaluating the board based on games played by human experts. Then, the scores of each node are re-adjusted by the score of their children nodes. Deep Blue choose the next move with the highest score in this tree. In comparison, AlphaGo (Silver et al., 2016) does all of the above as well. It runs a Monte Carlo Tree Search on the ensuing possibilities, uses supervised learning based on human expert games to generate policy, refines the policy by self-playing and reinforcement learning, and uses the policy to evaluate the board and find the next move with the highest score in the tree. In this respect, some may see AlphaGo as an improvement of Deep Blue with some new mechanisms.

However, AlphaGo can only be developed with the help of another development approach. While Deep Blue is an expert system developed completely under the *symbolic approach*, AlphaGo, for making policy, uses a deep neural network developed according to the *statistical approach*. Applying neural network in the development not only overcomes many problems that were deemed unsolvable (Johnson, 1997) at the time of Deep Blue, such as the explosion in search possibilities and the vagueness of the evaluation of the board by an expert but makes AlphaGo a mixed product of the two approaches in AI.

There is a consensus among Ai-theorists that the symbolic approach is distinct from the statistical approach in the field of AI. The two approaches have a different take on developing AI. The symbolic approach is often called a top-down approach, as it tries to emulate high-level human reasoning. The statistical approach is often called a bottom-up approach, as it tries to utilize the basic mechanism of data processing of humans. The two approaches produce models with different technologies. While the symbolic approach develops models with the representation of knowledge, logical inferential structure and search techniques, the statistical approach develops regression, classification and clustering using various techniques such as support vector machine, decision tree, Bayesian networks, and deep neural network. The two approaches rely on different resources to be functional. The symbolic approach requires a large amount of human expertise to solve problems in a specific area. The statistical approach requires a large volume of clean data and computational power in training. The two approaches also have different problems. The symbolic models of Ai are accused of being rigid, while the statistical models are accused of being transparent.

Because AlphaGo is partially modelled from the symbolic approach, which is distinct from the statistical approach. It seems inappropriate to call AlphaGo an enhancement of Deep Blue. In general, the development of models of Ai can not be explained only by the enhancement of models. For example, even within the statistical approach, each technique seems distinct from the others. Reinforcement learning and generative adversarial learning are two of the most popular approaches in recent years. But if we look at the mechanisms of reinforcement learning and generative adversarial learning, they are again two distinct approaches with different focuses, techniques, and problems. While reinforcement learning focus on reinforcing the net over policy, generative adversarial learning focuses on learning with the competition. They can both find effective use scenarios independently or together. It is inappropriate to call one the enhancement of the other. The enhancement view can not explain these new developments.

4.1.2. Developing relevant features for AI models

The more appropriate view of development is the featuring view of development. Under the view, development of models of AI consists of discovering new features of intelligence and collecting these features to make a model of AI. Enhancement of these features and the ways these features come together should be done only at the later stage. This view can explain the actual development of models of AI better. In the second half of the 20th centaury, AI-theorists have been focusing on developing features of knowledge, logical inferences, and careful searching under the symbolic approach. Moving into 21th centaury, AI-theorists understand how to develop features for pattern-recognition under the statistical approach. Such path of development is not a steady enhancement of a single model, but intermittent waves of feature addition and structural renewal.

Under the featuring view, it is important to decide if a new feature is relevant or it is not. If a feature is relevant, then we will attempt to add the new feature into our model. Intuitively, a feature is relevant if it is a crucial aspect of intelligence. For example, knowledge, searching, and logic are related to how human intelligence works. That s why theorists in the symbolic approach find them relevant. Pattern-recognition is how recognize objects in perception and words in speech. Hence they are relevant features for modelling AI. But as WCRA theorists reject the explicit reference to intelligence in the weak thesis. They can not appeal to the relation of these features to intelligence to decide if they are relevant.

WCRA theorists may argue that we should determine the relevancy of the features by testing how much performance they will add to a model if we add this feature in. But this way of deciding the relevancy of new features could run into two kinds of problems. First, some features only work if the model has certain architecture and features to cooperate with them. For example, artificial neural network in the statistical approach has been side-lined at first because we do not have sufficient data and computation power to see the improvement in performance. But moving into the 21st centaury, AI-theorists realize the potential it has with computational power and additional resources, it becomes the focus point again. If we use the weak in 1960, we would have judge the relevancy of deep learning to AI wrong.

Second, one may run into the danger of focusing on minor improvement only. Some not so relevant factors may bring some limited improvement quickly. But continue enhancing these factors can not bring further advancement. Here, I am not trying to argue that we should not take time to develop some quick way of enhancing the model if its benefit is immediate. But the addition of these thesis may not move us closer to what we want in the long run. AI-theorists have experienced these kind of incidents in periods of AI winters, when they continue to progress but stop producing break-throughs. Many of the new techniques that help Ai-theorists out of the Ai winters may have been overshadowed by the immediate performance increase of the minor improvement.

Overall, it seems instrumental rationality can not tell what is relevant and it has not told AI-theorists what is relevant in their development. This seems to be good evident why the weak thesis is wrong. We should posit intelligence explicitly in our pursuit to decide what is relevant.

4.1.3. The need for intelligence as the development goal

At last, the inability of instrumental rationality to tell what is a relevant development approach shows the implausibility of the emergence thesis. Emergence claims that through developing instrumentally rational machine, intelligence will emerge as a property of the machine. This seems to unlikely. If we look at the past developments of AI, it seems more likely that if we do not add the relevant features to the machine, the machines will lack that features no matter how well it performs. For instance, if we do not use neural network in the development of AlphaGo, then it can not recognize patterns in the board and evaluate them. We do not expect by give a Go version of Deep Blue more computation power and expertise it needs to play Go, it will emerge to recognize patterns on the board. Even if a Go version of Deep Blue manage to succeed in playing Go, it does not succeed by emerging to recognize the patterns on the board. The emergence thesis of intelligence is simply unsupported.

Here, I would like to contrast emergence thesis of intelligence with the converse thesis of emergence that is more plausible. Some AI-theorists appear to support emergece thesis, but in fact they do not. For example, Ai-theorists like Downing (Downing, 2015) conceptualize intelligence as instrumental rationality due to their endorsement of the strong thesis in the first place. Then they argue that rationality as intelligence is unimportant. They then claim that we should focus on the relevant mechanisms for intelligence that allows ‘intelligence’ defined by instrumental rationality to emerge. But as an agent gains instrumental rationality by having the relevant mechanisms for intelligence, these theorists are actually argue for an emergence thesis in converse. In this emergence thesis in converse, they mean that instrumental rationality can emerge from intelligence. This emergence is consistent with what AI-theorists have been doing. If this is the case, then it shows that these theorists do not support the emergence of intelligence from rationality. They also do not support the weak thesis, as they have been explicitly positing the mechanism of intelligence as their goal of pursuit.

4.2. Problems in application

This sub-section 4.2, we will discuss problems with the rational agency approach in applications. In applications, we care more than rationality. we care if the machines we deploy are beneficial and want to know that the machines are beneficial. To obtain knowledge of machines' beneficialness, I will argue that we need to pursue intelligence overtly. To show this, I will argue that it is difficult to know the beneficialness of unintelligent machines.

4.2.1. The need for beneficialness

In the book *human compatible: Ai and the problem of control* (Russell, 2019), Russell describes the danger of applying instrumentally rational machines. He argues that machines performing well in achieving our end goals may fail us. Russell uses the myth of King Midas to demonstrate this problem. In the myth, King Midas was 'blessed' with the ability to turn everything he touched into gold, which seemed to satisfy his craving for gold. But Midas failed to envision that he would turn any person, food and drink into gold upon his contact and ended up starving miserably to death. Now, suppose someone who shares the same desire as King Midas asks a capable and rational machine to exchange everything he touches for an equal mass of gold. Again, the machine endorsing this end goal may end up starving the client to death while fulfilling his gold. Although this case is likely exaggerated, it shows the possibility that machines can be rational without being beneficial.

Problems like this happen when rational machines endorse wrong instrumental sub-goals for the end goals. An instrumental sub-goal to the main goal is a sub-goal that can advance the achievement of the main goal. In cases like king Midas, we fail to recognize the badness of an instrumental sub-goal and dictate wrong goals to machines, which causes tragedy. In addition, humans often fail to recognize bad but extremely effective *instrumental sub-goals* (2019, p141) to their end goals, while AI can recognize them and endorse the wrong sub-goals for their effectiveness. In recent years, problems like such have been brought to light with the applications of AI. For example, many recommender systems may tend to promote evocative but non-educative content, including extremism and eroticism, which can be an effective instrumental sub-goal to the main goal of recommending content fitting the user's personal taste because everyone can be easily promoted to be attracted to evocative content. If their taste changes to fit the evocative content, then the main goal of fitting their taste will be achieved effectively. However, promoting universally evocative content is morally wrong, which has caused serious problems in our society.

In response, Russell argues that we need to make sure that we only apply models of AI achieving goals in line with our actual preferences. For example, the client's goal to exchange everything he touches as an equal mass of gold shows his preference for gold over many other things. However, if the machine learns the client as a person, the machine will learn that the client nonetheless prefers living over dying with a lot of gold. Then a beneficial machine will act only according to the actual preference of King Midas, which is to give him as much gold as possible without making him unable to live. For example, although we want to view engaging content recommended by the recommender systems, we prefer to view educative content much more than purely evocative content. If only evocative content is available, we may prefer not to avoid viewing them. Then beneficial machine which learns our actual preferences will not promote evocative content to increase engagement.

In some simple cases, our preferences may be easily learnt by AI. (2019, p192-203). However, our preferences can be extremely complex and hard to learn. Our preferences are often closely linked to our reasons and abstract standard. When we have different reasons for the end goals, we prefer different ways of doing things. For example, Russell (2019, Sec 8.4) raises the question about what a beneficial machine should do upon requesting a coffee for a user when there is no coffee nearby. If the user is merely thirsty, she does not mind getting a cup of water. If the user has an important meeting in 30 minutes but feels tired, she needs other energy drinks. However, as discussed in Sections 3.3 and 3.4, instrumentally rational machines may not be able to identify causal relations and act robustly according to our abstract standard. But it is almost impossible for the machine to see this preference without grasping the causal relation between the reasons for the end goals and the preferred ways of achieving them. Then it is difficult for an unintelligent rational machine to learn and act according to our preferences. To be beneficial, we should make these machines intelligent in the first place.

4.2.2. The need for active explain-ability

Perhaps, there are sufficient patterns in human actions. Then without learning our reasons for those preferences in the circumstances, machines can directly establish an association between our preferences with the corresponding circumstances. But explanations of machines' working mechanism are also important. As I will argue in the following sub-section 4.2.2, I will point to two contexts of applications where active explanations for the machines are important. First, in jurisdictions, the rule of law is a fundamental value pertinent to the autonomy of individuals. I will argue that explanations are necessary for the rule of law. In education, truth and understanding have special values in our lives. Without truthful explanations, we will not be satisfied by machines' applications.

In jurisdictions, respect for the rules of law is crucial. Based on autonomy, Tasioulas (Tasioulas, 2023) argues why respect for the rules of law is important and why explanation is crucial. The rule of law is a value encompassing several principles governing legal decision-making. These principles include making a decision based on general and formal laws only, making clear and non-contradictory laws, and that effective laws should be promulgated in advance. (2023, p4). Tasioulas argues that these principles allow the human beings, on behalf of whom the machine is making a decision, 'to grasp and factor into their rational deliberation in advance of their own decision-making (2023, p13)', which, Tasioulas argues, respects our autonomy to evaluate the situation and decide to follow the decisions of the machines in the particular case. (2023, p11). Only if the decisions of machines are made under the rules of laws can human beings comply with the decision of machines without relinquishing the dignity to act autonomously. But without an explanation of how the machines' decisions are made, we can not know if the machine arrives at the judgement according to the general and formal laws; we can not know if the machine arrives at the decisions with a clear and non-contradictory process. Hence, if machines can not actively explain its decision, their decision does not adhere to the rule of law and violates human dignity to be autonomous.

Another reason why explanation is crucial is that we as beneficiary care about truths. To demonstrate this, suppose we have a machine that has figured out how to perfectly serve human beings. It can act according to what we need in any circumstance and ensure that we are fed and entertained. Whenever we need to eat something, the machine will hand over food to us. If you are bored, the machine can find the most suitable pieces of music for you. But we will not consider this machine as truly beneficial because we do not only want to be fed or entertained. We want to understand how and why we are fed and entertained. We want to find out truths about our life and the world. We want to understand the truths behind machines' decisions. But to teach us the truth, a machine must be able to explain its decisions actively and truthfully. This requirement is more than satisfyingly answer a question about its action, like the current Chatbots do. It is not hard to imagine that a Chatbot can satisfying answer such questions by fooling us with answers we will likely give to our actions. To tell the truth that we can learn from it, it requires the machine to learn the robust casual relations their statements, which is unlikely to be achieved unless they are intelligent.

Therefore, active[18] explainablity from the machine is crucial for an ethical application of AI, and active explainability is best realized on intelligent machines. An intelligent machine can learn the preferences of human beings by picking up our knowledge and asking for our reasons behind ends. They can explain to us what their decisions are and respect the rules of laws and our autonomy. They can educate us about how something can be done and help us understand the world. Since we have these desiderata in applications, it further supports the view that we should make intelligence the explicit pursuit in the field of AI. These problems in application show further that AIs are not just instrumentally rational machines, and it is important for us to develop AIs that are intelligent.

[18] Active explainability should be distinguished from the passive explainability where we give explanation for the machine. Passive explainability is not satisfying. In future applications, when machine do things that are too complicated for us to comprehend in the first place. We need the machine to explain things to us actively.

5. An alternative approach: motivation adaptation approach

5.1. Defining motivation adaptation approach

In the introduction, we have made the distinction between the desire-based account of an agent and the value-based account of an agent. In sub-section 2.1, we argued that the rational agency approach is a natural development from the desire-based account of the agent. In this sub-section 5.1, I will first sharpen the distinction between the two accounts of an agent. Then, I will develop the value-based account of an agent and define the intelligence upon the value-based account of intelligence.

5.1.1. Value-based account of agent

To start, I will explain two common elements in both accounts of an agent. First, minimally, we conceive that an agent has the ability to initiate actions. This minimal condition means an agent possesses something that allows it to change itself or the world under its control. This something can be called motivation. While action is individuated by the changes it initiates, *motivation* is individuated by the activation condition, an expectation of action after its activation and the process of generating action. For example, when an agent engages in grabbing an apple. It has the motivation to grab an apple, which consists of the expectation to grab the apple, the condition of engaging the grab (i.e. hunger), and an initiating process to generate the action of grabbing the apple. For initiation, the motivation can activate other related motivations, such as the motivations for the agent to round its hand, grip and lift the apple. Motivation is crucial to the conception of an agent. Hence, in both accounts of agents, motivations are included.

The two accounts of agents diverge in relating motivations and actions. In the desire-based account of an agent, it is a *teleological relation* between motivations and actions. According to the teleological relation, motivations exist for actions. The goodness of motivation is judged by the goodness of the action it expects to initiate. The best action is one that brings the desire of the agents. In this regard, desires form the fixed standard used to judge the performance of actions and the goodness of motivations. As I have discussed in section 2.1, rationality is a natural development from the teleological relation. In addition, many theories of well-being also lay their foundation on the teleological relation between motivations and actions. For example, according to the objective-list theory of well-being, an agent is living well only if the events in the list of well-being obtain. Then a good set of motivations for well-being will be a set of motivations that can initiate events of well-being.

In comparison, according to the value-based account of an agent, the relation between the motivations and actions is *generative*. In this account, an action is generated by motivation. Under the generation, we judge the goodness of actions by the motivations behind the action. In turn, we judge the goodness of motivations by the values they stand for. For example, pleasure and pain can give value to motivations. However, while pleasure is a good value, pain is a bad value. For example, a student can be motivated by fear of reproach to finish the assignment, or the student can be motivated by a passion for knowledge. Because the former motivation stands for pain and the latter for pleasure, the former action is better even if the two motivations generate the same actions.

Here, the goal is not to adjudicate between the account of agents. Both accounts of an agent can co-exist at the same time, as both generative and teleological relations can present between motivations and actions. Instead, the goal is to take the value-based account as a foundation to explore the concept of intelligence. Establishing a conception of intelligence based on the value-based account is reasonable. This is because questions about intelligent action and agents are questions about the generations of action rather than the purpose of actions. This focus on generation is supported by the etymology of 'intelligence'. The word 'intelligence' comes from Latin 'intelligere', which means to understand. Intuitively, to understand something is to use a mechanism to form motivations behind actions; to understand is not to have some particular purpose for actions or to fulfil a desire for action. Following this line of thought, it should be wrong to theorize intelligence from the perspective of the desire-based account of the agent.

Finally, according to the value-based account of an agent, a generative motivation is conceived as a unit that generates action with expectation under conditions. As our discussions will be based on the value-based account of an agent, I will henceforth refer to motivation as generative motivation exclusively

5.1.2. Concieving intelligence by adapt-ability

According to the value-based account of an agent, a good motivation can stand for values of goodness for the agent. Hence, one must find the value associated with intelligence to find a conception of intelligent motivations. Intuitively, the value of intelligence is the realization of control. Therefore, an intelligent agent can have motivations for the agent to control its action. Hence, the first conceptual task for us is to specify the concept of control. Under the value-based account of an agent, it is natural to see the coincidence of expectation and generated action as the indication of control. If an agent acts in what is expected, it is natural to say that the agent controls its action. Hence, following this line of thought, we can first specify what it means to have a *controllable motivation* in the following way: A controllable motivation is one that generates action expected under the conditions.

However, it is not the case that an agent is more intelligent if it has only controllable motivations. In the extreme case, an agent may have only one motivation with null expectation and null generation. In this case, the agent has only controllable motivation but is as static as a stone and very far from intelligent. In practice, an agent is expected to have uncontrollable motivations. This can happen when the expectation of a motivation is fundamentally impossible to be achieved. In general, it is impossible to expect anything that violates mathematics and the laws of physics. For instance, a motivation can not expect to hit a wall without being hit by the wall, which violates Newton's third law. In other cases, a motivation can be uncontrollable when the generation is wrong. An example will be singing a note C with the wrong frequency. In this case, the expectation can be achieved, but the generation of the expectation needs to be corrected. Compared to possessing controllable motivations, the ability to increase the numbers and proportions of controllable motivations is more important. This ability can be specified as *motivation adaptation*. One can define motivation adaptation as the mechanism that allows the addition and adjustment of motivations for control. Motivation adaptation can occur in many ways. It can happen by changing the expectation of the motivation to meet the actual generation of the motivation. It can also happen by changing the process of generations to meet expectations. In other cases, it can happen in exploration to form new expectations and generations for the expectation.

In the following, I will give a complete example of motivation adaptation. Suppose one expects to sing the pitch C but generates a frequency of vibration in the vocal chord to hit pitch B. In this case, an uncontrollable motivation occurs. The first step of adaptation may be to change the expectation of the generation to identify what pitch is produced. By changing the expectation to pitch B, he can identify the frequency for pitch B. Now he may create a new motivation with a new way of generating vibration in a shortened vocal chord with a higher frequency. In the process, he constantly adjusts the frequency until the expectation is met. When the expected pitch C frequency is found, the agent may practice singing the pitch to consolidate the creation of the new motivation. He may also try to sing the two pitch side by side to differentiate them. After this process, the agent adapts to controllable motivations for singing both pitches B and C.

As a notice, motivation adaptation should be distinguished from some conception of adaptation in the literature under which what is changed is not generative motivation. We may call such conception *instrumental adaptation*. For example, genetic evolution is an instrumental adaptation but not a motivation adaptation. Genes are not generative motivations. If we look at complex multicellular organisms, we do not think gene generates or expects particular actions of the agent directly. Although gene encodes information on how the organism should be built and developed, they do not directly participate in generating the agent's action. In addition, genes normally do not adjust in order to meet actions with expectations. Empirical evidence suggests that genes are relatively fixed throughout the organism's lifetime, even when many unexpected actions happen in the meantime. Genes are better seen as part of the physical body of the organism. In this sense, evolution is a process that changes the agent's physical body without changing the generative motivations. Hence, evolution is an instrumental adaptation but not a motivation adaptation. For the convenience of discussion, I shall henceforth assume the exclusivity of adaptation as motivation adaptation.[19]

Finally, under the account of the value-based agent with the concepts of controllable motivations and adaptation, one can define intelligence as the following: an agent is intelligent

[19] Similarly, it is debateable if a physical neuron is a motivation, as individual neuron does not seem to expect actions. Physically, a motivation might be a collection of neurons with some structure corresponding to knowledge/reason/belief/desire. However, the determination of the physical realizers of a motivation is up to the AI theorists. It is beyond the scope of this thesis to expound on this question.

if it has the ability to adapt.[20] In other words, an agent is intelligent only if it has the mechanism to add and adjust motivations when the expectations of the motivations do not meet the action generated by the motivations. Agent intelligence is the primary sense of intelligence. In derivative, one can define action intelligence by narrowing the circumstances under which the ability to adapt is relevant: an action is intelligent if the agent can adapt motivations when the action and the consequences of the action are unexpected. In many ways, this definition is intuitive. Intelligent human beings are adaptable. We can use our intelligence to adapt to the situation when things go unexpected. Factors in the IQ test are chosen because they are crucial to a human being's ability to adapt. For example, according to Binet, one of the founders of the study of IQ tests, intelligence should be conceived as 'judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances.' (Binet & Simon, 1916, pp. 42-43).[21]

To conclude this section, I have demonstrated the basic conceptual framework for the motivational adaptation approach. It is built from the value-based account of an agent, with concepts of control and adaptation. Under this framework, the motivational adaptation approach can define intelligence as having the ability to adapt. Similar to the rational agency approach, to put the motivation adaptation approach into work, one first specifies more concepts to complete the description of the adaptation mechanism in the conception phase. Then, AI theorists will try to develop features that can realize the mechanism of adaptation on a value-based agent in the development phase. At last, AI theorists will adjust the development in the face of the challenges in the application phase. In the next section, I will suggest how this approach works in practice compared to the rational agency approach. However, completing this project will be out of this essay's scope. The focus of the next section will be to argue that the motivation adaptation approach is better than the rational agency approach on questions we raised in sections 3 and 4.

[20] Another important definition of intelligence is in the comparative sense. One can conceive of this definition. In the comparative sense, an agent X is more intelligent than the other agent if the agent's ability to adapt is better than the other agent.

[21] This sense of adaptation also explains the etymology of intelligence. For example, in Heidegger's analysis of understanding, an agent has a network of affordances (Wrathall, 2013)**.** Hence, according to Heidegger, to understand something is to find the affordance for the thing. As an affordance involves both a generation process and expectation from the way it is understood, we can conceptually inject the concept of motivation homomorphically as the concept of affordance in Heidegger's analysis of understanding. Then intelligence, as the ability to understand, will be the ability to adapt. This sketchy argument shows how the motivation adaptation approach preserves the connection between understanding, adaptation and intelligence. Unfortunately, it is beyond the scope of this thesis to expound this connection.]

5.2. Answering the problems against rational agency approach

In this sub-section 5.2, I will answer problems raised against the rational agency approach by the motivation adaptation approach. Sections 5.2.1, 5.2.2, and 5.2.3 will answer problems in conception, development and application respectively in this order.

5.2.1. answering the problems in conception

To answer the problem of cognitive slack, one can see that intelligence as having adaptability can answer the counter-example effectively. In the counter-example of cognitive slack, Peter can find the hypothenuse of a right triangle by drawing without using the Pythagorean theorem. This discovery shows that Peter has good adaptability. Initially, Peter does not have the motivation to generate the answer about the hypothenuse of the right triangle. But then he finds a way to meet this expectation by finding drawing and measuring as a method to generate the right triangle. Later, when he is taught the Pythagorean theorem, he is able to learn how to generate the answer and when (i.e. in the case of the right triangle) this answer can be applied. So it also shows the intelligence of Peter when learning the Pythagorean theorem. However, applying the Pythagorean after acquisition falls short of showing the adaptation process as Peter only acts from an existing motivation. This absence of new adaptation explains why we think applying a powerful yet blindly routine method can strike us as a process without intelligence.

In the counter-example of shallow success, we explain why the action of the vegetative investor and Mary is unintelligent but rational, but agent-wise, Mary is intelligent, and the vegetative investor is unintelligent. One can explain why this is the case by the motivation adaptation approach. I will explain their attribution of intelligence one by one. First, because the vegetative investor is unaware of what is happening in the real estate market, the vegetative investor can not adapt motivations and actions for new investment when the consequence of his current choice investment is unexpected. Because of this unexpectedness, the investor is not intelligent in his investment. Overall, the investor is unaware of everything; hence, he cannot adapt when the investment goes unexpected. Thus, the vegetative investor is an unintelligent agent. Second, because Mary only cares about her residence in the place, she does not make her investment decision based on the consequences of the investments. When the market changes, she does not adapt to the change. Hence, she is not intelligent in investment. However, Mary does adapt to changes in other scenarios. Her adaptation means that Mary is an intelligent agent.

In the counter-example in the problem of abstract robustness, we show that a Chatbot can answer a selection of questions satisfyingly while showing significant inconsistency and complete misunderstanding in cases where we expect it to do well. Intuitively, this shows that the Chatbot is not an intelligent agent. In the framework of the motivation adaptation approach, the case of the Chatbot shows that the Chatbot's adjustment of actions is not a process of adaptation that can coincide actions with expectations. According to its design, the expectation of the Chatbot is to answer the question, but the adjustment of the Chatbot's output is not made based on whether it answers the questions. On the contrary, the Chatbot is adjusted only in a way that it can produce answers similar to the answers to give an existing questions poll. This adjustment allows the Chatbot to produce satisfying answers to questions when similar questions and high-quality answers are available in the poll. However, these data-oriented adjustments are insufficient to adjust the Chatbot to meet the expectation to answer the questions. Indeed, with the increase in the amount of data in the poll and the complexity of techniques to make similar answers, the Chatbot like ChatGPT can answer a wide range of questions with high quality; however, as revealed by the case, we have about ChatGPT, the Chatbot, as such designed does not answer the question according to the semantics of the question. Because of this, these Chatbots' adjustments can not be called adaptation, as adaptation is the process of making actions and expectations coincide. Therefore, these Chatbots cannot adapt, and they are unintelligent.

The problem of causal learning points to the inability of current AI models to learn causally while being able to deliver good performance like a super-animal. In section 3.4, argue that this shows that the machine is unintelligent. This case can also be explained in the motivation adaptation approach. According to the value-based account of an agent under this approach, a machine has the ability to engage in motivational adaptation only if the machine has generative motivations. Individually, generative motivation is a unit for generation and expectation. The action of motivation can be targeted at other motivations. A motivation can generate actions that can activate other actions, be a condition of other motivations, or change the generation and expectation of other motivations in the adaptation mechanism. Hence, motivation can form a network in which we take motivations as nodes and their interactions as arcs. In this respect, a network of motivation is similar to the artificial neural network widely applied in recent years.

But as the current neural network fails to do causal learning, it shows that the artificial neuron as we design today falls short of being a motivation. The inability to do causal learning can also be inferred from how artificial neural networks are designed. An artificial neuron, also called a perceptron, generates parameters for other neurons to determine their generations. However, according to the regular design of neural networks, expectation only happens at the output layers. At the output layer, the whole neural network is put under the test of accuracy by a cost function, and adjustment is made to reduce the cost. This design means that each perceptron does not expect anything and falls short of the generative relation. Because of the lack of expectation in the generation of the final output action, the generation can not be broken down into a series of expectations leading to the actions.

A series of expectations can make a causal analysis of the final output action. For example, suppose one drinks water from a water bottle. There is then a series of motivations to be activated. He is motivated to lift the hand in expectation to lift the bottle he is holding; he is motivated to lift the bottle he is holding in expecting to deliver the water in the bottle closer to his mouth. In this analysis, it is conceptually reasonable to say that each step is causal. In the example, it is reasonable to say that lifting his hand caused the water bottle to move closer to his mouth. This example shows that a motivational structure is sufficient to deliver a causal analysis, but a neural network is insufficient to deliver a causal analysis by its current design. Because of the absence of a generative structure of motivation for its current design, the AI models based on the current technology of neural networks can not have the ability to make motivation adaptations. Hence, this explains why the current AI models are unintelligent despite their high performance.

5.2.2. reply to problems in developments

In the development phase of the motivation adaptation approach, the goal is to make a machine able to make motivation adaptation. This development can be further broken down into two aspects. First, developing motivation adaptation would require AI theorists to design a unit that can realise generative motivation and structure the motivations into a network. Second, this development requires the development of an adaptation mechanism on top of the network of motivations. This development may be done by giving some properties to a motivation that can make it responsive to the difference in expectation and generation. Alternatively, the adaptation might be reduced to a core of basic motivations that can generate the adaptation process in the agent when the basic motivations are activated. The direction of adaptation can also be affected by the existing motivations. For example, if one knows how to increase or decrease one's pitch and how to sing pitch B, then one can learn to sing pitch C by increasing the frequency on the basis of pitch B. These two aspects must be developed to find true AI.

Realistically, this would require many attempts to try different models of AI with different features until the AI is developed. This trial and error process can explain the past developments of AI models. Roughly speaking, the conditional statements in the expert systems can be seen as an early attempt to model motivations. However, conditional statements have a tough time generating detailed actions as the output is qualitative, and many motor actions and perceptual characterisations of robots require quantitative output. Another problem is that the adjustment mechanism is too limited in conditional statements. The addition of new conditional statements must be the combination of elements specified manually by the developer, and the adjustment only makes big jumps within these manually specified elements. Later in the development, perceptron comes into the spotlight as a model for motivation. In many ways, the perceptron is a quantitative conditional statement, as the output number of the neuron is conditional on the input number. Adding this feature of quantitative change allows the AI model to generate a wider class of actions and engage in more subtle changes than it previously could. But as we have discussed, the additional featuring of quantitative change is insufficient to make neural networks, generally, a model for motivations because it lacks the design of expectation. Many current developments in the field of AI can be seen as an attempt to solve this problem. For example, the LSTM network uses a mechanism that allows the basic unit to store more information. The basic unit, which is often called a cell, could be able to store information that models the expectation of a cell.

The attempt to find the adaptation mechanism can also explain trends like multi-modal and cross-modal training of machines. At the moment, Chatbot models cannot generate answers that can meet their expectations to understand the meaning. Take ChatGPT as an example; it is only able to generate answers based on the syntactical significance of the sentence but not semantics. In comparison, when human beings say a triangle is sitting above a square, we are not only generating the sequence of words' a triangle is sitting above a square', but we also have a vivid picture of what it is like for a triangle to sit above a square. ChatGPT is expected to do the same, but it is never given the means to do so, as it is not trained to generate pictures with only linguistic data. In this sense, ChatGPT is destined to fail to generate what it is expected to. To better model the process of adaptation, machines should be able to adjust under the input and output of a full range of conditions as humans can. The cross-modal development of AI models is another trend that is hard to be explained only from the perspective of the rational agency approach.

Motivation adaptation provides a much clearer goal for developers to achieve than rationality. One can see the relevancy of a potential feature through its design. In addition, intuition on how adaptation works can help developers come up with relevant features in the field. However, as a disclaimer, I am not arguing that one should completely disregard rationality as one of the pursuits of Ai-theorists. For application, one can develop AI models with good performance without adding more features for rapid development. Nonetheless, to develop true AI, one must develop features for realising motivation adaptation.

Moreover, rationality can emerge in the development of machines that can adapt. As the machines have better models of motivation adaptation, by conception, they will become better at adjusting its action to be controllable in a wide range of scenarios. Then naturally, the model's performance on the tasks it adapts to will become high. In order to increase the rationality of the machine, one should also focus on developing motivation adaptation in machines.

5.2.3. reply to problems in applications

We raise two concerns against the rational agency approach in the problems in application. First, there is the worry that machines will endorse catastrophic instrumental sub-goals for a specific task. Second, there is the worry that machines can not explain and teach us the sub-goals necessary to endorse, which undermines many crucial applications, like jurisdiction and education. For these two concerns, It is unclear how they can be alleviated by enhancing rationality. In comparison, the two concerns can be more easily digested if a machine has the structure of motivations and the adaptation mechanism.

To start, I will discuss how a machine can be beneficial according to the motivation adaptation approach. In general, the action of the machine is good if a network of good motivations generates it. The values behind good motivations can be judged according to various ethical theories. For example, according to Kantians, good motivation is universalisable as a law. According to utilitarianism, good motivation brings the most happiness in action. When motivated to act according to ethical theories, these theories will provide conditions for generating actions. Hence, values held by the Kantians and the utilitarianists construed as such are high-level motivations acting on the low-level motivations for external actions. Because of this, an agent can endorse an ethical value by adapting high-level motivations for the ethical theory.

When the agent endorses an ethical value, it will modify the adaptation process of motivations. It will limit the range of expectations of motivations. For example, utilitarianism motivates one to not adjust expectations in a way that will not bring maximal happiness for everybody. So, when an agent who endorsed utilitarianism unexpectedly does something detrimental to others but benefits greatly, the agent can only adapt to this motivation by changing the generation process. Therefore, morality can be learnt in the process of value-based adaptation. Because moral values as high-level motivations can participate in the adaptation process, this will preclude the possibility of endorsing a wrong instrumental sub-goal for a good end goal. For example, a moral agent will not develop motivations that expect to give someone the ability to turn everything into gold because this action, judged individually, is seriously dangerous. Hence even when the agent is given the task to satisfy the wishes of King Midas as much as possible, it will not endow the King with the ability to turn everything into gold because it lacks the low-level motivations to generate the unethical action. In general, under the morally guided value-based adaptation, an AI will not develop wrong motivations, so it will not generate actions that are wrong but useful only for a specific task. This avoidance in creating explains how the concern about endorsing wrong instrumental sub-goals can be solved by developing an ethically guided adaptation mechanism.

For the concern about active explainability, it is natural to see the active explanation as an extension of the motivation adaptation mechanism. In active explanation, the primary purpose is to convey motivation from one party to the other. That is to say, active explanation is a mechanism that can generate a message for motivations that allow the recipients to adapt to the motivation behind the message. This mechanism can be further broken down into two related adaptation processes. First, there is the process of adapting to the language to generate the message. Second, there is the process of adapting to the motivation behind the message. Unfortunately, it is beyond the scope of this essay to study the mechanism of these two processes. Nevertheless, as we have discussed, the realisation of active explainability is based on the pre-conditions of a machine having a motivation structure and the ability to adapt. Again, it is unlikely that active explainability can be developed if AI theorists only focus on enhancing rationality. More plausibly, Ai theorists should look for features relevant to the process of active explanation and include them in their AI models. Therefore, from the perspective of developing active explainability on machines, it is crucial to develop motivation adaptation. The foundational role of motivation adaptation for active explainability further supports the motivation adaptation approach to AI.

6. Conclusion

In this thesis, I have argued against the claim that AI is intelligent because of instrumental rationality. Specifically, I raise problems against the rational agency approach in its conception, development and application. Conceptually, one can not find a specification of instrumental rationality such that intelligence can be reduced to it. For any specifications, there are counterexamples against the threshold or the selection of test environments. Because of these counterexamples, one can not argue for the claim that AI is intelligent because of instrumental rationality by the reduction thesis. Another way to argue for the thesis without a reduction thesis is by the emergence thesis. According to the emergence thesis, intelligence can emerge from machines as we develop better and better instrumentally rational machines. In response, I argue against the emergence thesis by the problems in development. I argue that in the actual development process of AI models, AI theorists require intelligence as the explicit goal to find relevant features. Hence, Ai is neither intelligent because of instrumental rationality by the emergence thesis. In addition, wide applications of AI can only be safe and fair based on features developed on its intelligence. This requirement in application further supports the claim that an unintelligent machine is not ultimately what we want. Therefore, it is a considerable problem for the rational agency approach to fail to make machine intelligence.

Some theorists may complain that there is no better alternative approach than the rational agency approach, so we should pretend that AI can be intelligent because of instrumental rationality, even if it is false. To argue against this view, I propose a solid alternative named the motivation adaptation approach. Under this approach, a machine is intelligent only if it has a network of generative motivations and the ability to adapt the motivations. I have shown that this approach is intuitive conceptually from the perspective of a value-based account of the agent. I have shown how this account can avoid the conceptual counterexamples against the rational agency approach. Finally, I have shown how it can provide a clear goal for AI theorists in development and lay the foundation for ethical applications. All of these show that the motivation adaptation approach is better than the rational agency approach. As we can endorse the view that AI is intelligent because of its motivation adapt-ability, we can confidently conclude that AI is not intelligent because of instrumental rationality.

Bibliography

Altman, N. i Krzywinski, M. (2015). Association, correlation and causation. *Nat Methods 12*, 899–900. doi:https://doi.org/10.1038/nmeth.3587

Andreas, H. a. (2021). Regularity and Inferential Theories of Causation. Andreas, Holger and Mario Guenther, "Regularity and Inferential Theories of Causation", The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), . Pobrano z lokalizacji https://plato.stanford.edu/archives/fall2021/entries/causation-regularity/

Binet, A. i Simon, T. (1916). The intelligence of the feeble-minded. Baltimore, MD: Williams & Wilkins.

Buehner, M. J. i Cheng, P. W. (2005). Causal Learning. W K. J. Holyoak i R. G. Morrison (Redaktorzy), *The Cambridge handbook of thinking and reasoning* (strony 143-168). Cambridge University Press.

Campbell, M., Hoane, .. i Hsu, F.-h. (2002). Deep Blue. *Artificial Intelligence,* strony 57-83.

Dixit, A. (2014). *Microeconomics: A Very Short Introduction.* OUP Oxford.

Downing, K. L. (2015). *Intelligence Emerging Adaptivity and Search in Evolving Neural Systems.* MIT Press.

Grimm, S. (2021). *Understanding*. (E. N. Zalta, Redaktor) Pobrano z lokalizacji The Stanford Encyclopedia of Philosophy (Summer 2021 Edition): https://plato.stanford.edu/archives/sum2021/entries/understanding/

Haataja, M. i Bryson, J. J. (2019). Reflections on the EU AI act and how we could make it even better. *Antitrust Chronicle*.

Johnson, G. (1997). *To Test a Powerful Computer, Play an Ancient Game.* The New York Times.

Koralus, P. (2022). *Reason and Inquiry: The Erotetic Theory.* Oxford: Oxford University Press.

Lewis, D. K. (1973). *Counterfactuals.* Cambridge, MA, USA:: Blackwell.

McCarthy, J. (2007, Nov. 12). *What is artificial intelligence?* Pobrano z lokalizacji John McCarthy's Home Page: https://www-formal.stanford.edu/jmc/

Minsky, M. (1968). *Semantic Information Processing.* Cambridge: MIT Press.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* MIT Press.

Niko, Brunero, J i Kolodny. (2020). *Instrumental Rationality.* Pobrano z lokalizacji The Stanford Encyclopedia of Philosophy (Spring 2020 Edition): https://plato.stanford.edu/archives/spr2020/entries/rationality-instrumental/

Nobandegani, A., da Silva Castanheira, K., O'Donnell, T. i Shultz, T. (2019). On robustness: An undervalued dimension of human rationality. *Proceedings of ICCM 2019 - 17th International Conference on Cognitive Modelingpp.*, 126-131.

Pearl, J. (2019). The limitations of opaque learning machines . W J. B. (Ed.), *Possible Minds: 25 Ways of Looking at AI.* Penguin Press.

Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, 8.

Pinker, S. (2021). *Rationality: What It Is, Why It Seems Scarce, Why It Matters.* Penguin.

Russell, S. (2019). *Human Compatible: AI and the problem of control.* penguin random house.

Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist, 52(10)*, 1030–1037. doi:https://doi.org/10.1037/0003-066X.52.10.1030

Tasioulas, J. (2023). The Rule of Algorithm and the Rule of Law. *Vienna Lectures on Legal Philosophy (2023)*. Pobrano z lokalizacji https://ssrn.com/abstract=4319969 or http://dx.doi.org/10.2139/ssrn.4319969

Urquhart, A. (2004). Complexity. W L. Floridi (Red.), *The Blackwell Guide tothe philosophy of computing and information* . Blackwell Publishing Ltd.

Valiant, L. (2013). *Probably Approximately Correct.* Basic Books.

Weber, M. (1978). *Economy and Society.* (C. Wittich, Red., i g. roth, Tłum.) Berkeley: University of California Press.

Wrathall, M. (2013). Heidegger on Human Understanding . W M. Wrathall, *The cambridge companion to Heidegger's Being and Time* (strony 178-200). Cambridge University Press .