The Second-Person Standpoint and Moral Machine

Ma Hanlin [0000-0003-0589-3701]

Suzhou University of Science and Technology, Suzhou, China
fanfan2011cn2000@gmail.com

Abstract. Given the need for Human-Machine moral Interaction, an AI system that focuses on moral theory involving intersubjectivity is worthy of attention. We will build a network knowledge representation framework based on the second-person standpoint moral theory to analyze the design philosophy of a *general artificial intelligence* (AGI) moral machine. We then use this framework to investigate an interactive moral machine designed by Bello and Bringsjord (2013). In constructing a feature of mindreading, their study faces philosophical challenges to combine the empirical source configuration of its inheritance function and the multi-world setting of the *Polyscheme* itself. We try to prove that the integrated structure of the Polyscheme based on a third-person perspective, multi-world setting cannot afford a continuous self that can accommodate enough diversity of the stimulation relationship between the concept of self and other. Moreover, we suggest that the psychological distance structure of mindreading should mainly come from the internal "self" of the system. This requires the system to have an intrinsic kernel, meaning a design philosophy completely different from the integrated style of Polyscheme.

Keywords: second-person standpoint; moral machine; Artificial General Intelligence; human-machine interaction; AI system; mindreading

1. Introduction

With the improvement of research and application, the participation of AI in society is expected to increase significantly, and AI will increasingly enter into moral interactions. Therefore, researchers have steadily become concerned about how to make "machines" have ethical attributes. There are two types of moral machines. One functions only as an instrument for the moral enhancement of humans, while the other is a moral thinking machine with moral subjectivity. The former is a kind of specialized artificial intelligence technology designed to help humans obtain morally sound cognition, make moral decisions, comply with moral norms, and so on (Savulescu and Maslen 2015). Comprehensively introduced by Wendell Wallach and Colin Allen (2008), specialized AI is exemplified by Medical Ethics Expert (MedEthEx) designed by Michael Anderson and others (Anderson, Anderson, and Armen 2006), the Truth-Teller by McLaren and Ashley (1995), and Marcello Guarini's Ethical Case Classifier (the Moral Case Classifier, MCC in short) (Guarini

2011). The latter type of system is hoped to simultaneously realize the full range of human cognitive abilities and human-like moral attributes. For example, Wallach and Allen once collaborated with Stan Franklin, the creator of the LIDA(Learning Intelligent Distribution Agent) system, to explore how to implement a universal moral decision model relying on LIDA's high-level cognitive framework (Wallach, Franklin, and Allen 2010).

Moral machine scientists have their preferences for the mastery and application of moral theories. They focus on either specific applied ethical content (such as what MedEthEx does) or the normative properties of ethical norms itself (such as what MCC does). Others may be concerned with moral decision-making issues, referring to how to make moral systems emerge for particular collectives from systems representing conflicting moral imperatives (Dameski 2018). At present, it seems that only a few researchers consider AI moral systems from the perspective of inter-subject interactions. Therefore, moral-philosophical theories involving intersubjectivity are rarely utilized to assist in designing or evaluating a moral system. The second-person standpoint moral philosophy is dedicated to constructing moral knowledge that endorses intersubjectivity. It emphasizes decentralization, indeterminacy, and so forth. Influential ethical theories in the twentieth century, such as moral constructivism (Rawls 1980), virtue ethics (Williams 2008), and the ethics of care (Gilligan 1993) have potential connections with this theory. P. F. Strawson (2014), Stephen Darwall (2006), and R. Jay Wallace (2019) hold the most influential second-person theories of morality. Their central idea is that moral knowledge is modified in human interactions in the light of each other's reactions. Moral knowledge neither comes from the independent realm of the third-person nor is it purely the subjective construction of a first-person, but is determined by a multi-distributed network process. When building a second-person moral AGI system, we must thus consider how to establish a self-knowledge representation and control system compatible with these presuppositions and further establish a mindreading representation system involving intersubjectivity. From the standpoint of adopted design elements, the design philosophy of the specialized moral machine (SMM), the individual moral thinking machine (IMM), and the second-person moral thinking machine (S-pMM) can be compared through the following graph (Figure 1).

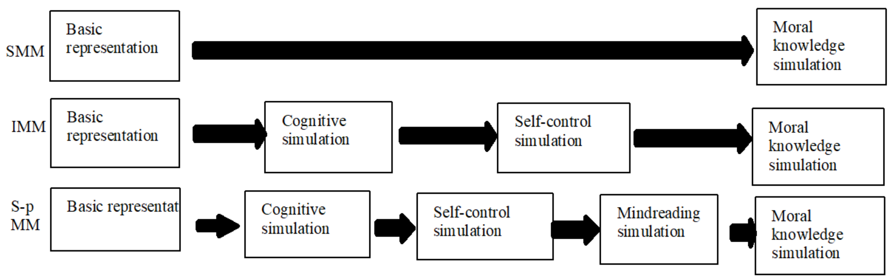

Fig. 1. Arrow indicates designing based on the former item

So far, only the work of Paul Bello and Selmer Bringsjord (2013) has involved the moral thinking machine construction from an intersubjective perspective. Their AGI model Polyscheme is one of the pioneers in the realm. Although the framework of Polyscheme they used has gradually declined, the traits of its design philosophy are still very representative, such as LIDA and OpenCog, which are still active in the same way as "integrating approach". Because of the framework's integrated basic structure and its reliance on possible-world semantics, its self-knowledge model will be faced with some difficulties, which blends with the problem encountered by Bello and Bringsjord when using a third-person criterion of inheritance function to construct intersubjectivity. The paper is organized as follows. In section 2, we present a second-person concept cycle network to analyze what the cognitive states should ascribe to. This network is joined by a *monotonic inheritance relationship*, which does not provide a *third-person* perspective to the computing process. In section 3, we introduce how Bello and Bringsjord characterize Knobe and Nichols's measurement of *causality involving self*, based on construal level theory (CLT). Next, in section 4, we will show that a central control of Polyscheme actually takes charge of the integrated box of possible worlds. In section 5, we find that this kind of *integrated* AGI system cannot build a self with continuous history, and cannot represent all the significant similarity between different possible worlds, which means different *selves*. In section 6, we introduce another AGI system (NARS) that may more closely match the second-person concept cycle network, and explain why this is so.

2. Second-person concept cycle network and intersubjectivity

Stephen Darwall systematically uses the second-person standpoint to study the issue of normative origin. In a nutshell, the *second-person cycle* he proposed refers to the fact that the subject and others in a sense share a way of *explaining* each other's behavior (*second-person reasoning*) and a way of *requiring* and *blaming* each other (requirement and accountability). This shared basis is called second-person authority, and all these concepts are mutually defined. For example, there is a connection between *requirement* and *accountability*, because people generally believe that moral *requirements* are based on *accountability* or *obligation*—if you don't do something you are supposed do, it means you fail to meet the requirements. Under the second-person theoretical presupposition, accountability is also in line with requirement, because the structure of accountability is based on Strawson's "*reactive attitude*." That is to say, I have to comply with certain responsibilities precisely because others have a reactive attitude toward me. If I step on someone else's feet, she/he will express resentment against me, which can be understood as expressing some kind of moral requirement and moral authority. It is based on this attitude that I establish the viewpoint of not stepping on others' feet. However, for a rational responsibility construction process, we have to look at the "reason" behind the attitude of others, which in turn is related to responsibility, because she/he will feel that you are responsible for this attitude. All of the above is connected to "second-person authority"—all people in the relationship believe that everyone recognizes the legitimacy of each other's reaction and attitude.

Formally speaking, the definition of the second-person normative concept is distributed, which means that no concept in the circle can be defined as an atom, thus determining the connotation and extension of other concepts. Therefore, we recommend using multiple inheritance relationships to describe this normativity. For example, the semantics of shared accountability by others and me can be inherited by the meaning of his requirement of me (which I'm also certain of), which is "you have to do it" (to be responsible to him) ; it is also inherited by reasons for others to condemn me morally (which I also agree with) to some extent; it is inherited by the second-person authority shared by my Self and others' Selves.

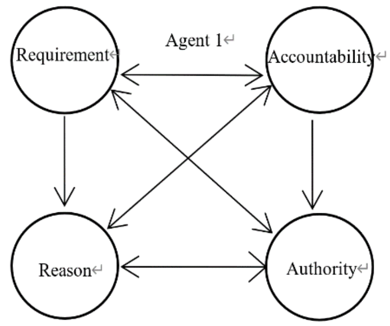

Fig. 2. Solid line indicates monotone inheritance

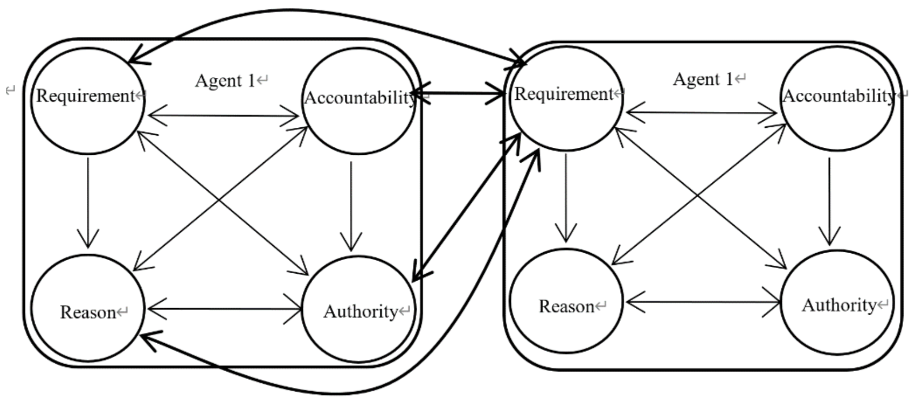

Fig. 3. Schematic diagram of a conceptual network between two agents in an ideal situation

For Darwall, it is also feasible to describe these concepts using inheritance relationships in other directions under ideal circumstances. From the perspective of knowledge representation, it is thus merely necessary to construct a model concerning one of the concepts, with the other three being partially represented. In an idealized situation, this is not a typical *inheritance logic*, because inheritance relationships are generally asymmetric (Al-Asady 1995). If the relationship between subjects is not involved, it can be assumed that the representation of the moral concept network is as shown in Figure 2. For example, the definition of the "reason" is determined by the relationship between it and other concepts. It is not difficult to implement a network alone. For instance, an artificial neural network (ANN) can theoretically meet the needs of this data structure. Apart from the distributed definition, the second-person perspective also requires the conceptual loop network to be characterized by a reactive attitude, namely the subject's judgment of other subjects' experiences. Here, a representation framework involving the subject (or the Self) is needed. How to design this framework and combine it with specific distributed technology paths will be the core task of building S-pMM.

The biggest challenge here is the contradiction between the moral knowledge representation and the computing knowledge representation. Darwall's *Kantian* tendency presupposed substantial completeness of his moral knowledge representation, which perceived norms as a *monotonic inheritance relationship* (Figure 2). Supposing we have acquired a self-representation framework for knowledge in some way and stayed with this framework, it is impossible for the machine to halt in the process of characterizing the second-person normative knowledge. As the representation of self-knowledge is first-person, the iterative subjective perspectives will explode in combination when some people assume that others' subjective perspectives are not equal to theirs. Take the rule "don't do to others what you don't want done to yourself" as an example. From the second-person perspective, this sentence can be reduced to the conjunction of the following two sentences: A = "We are unwilling to be the cause of situations we do not want to see" and B = "In our eyes, we share this causal mechanism with others." Consequently, this rule has become a normative law, and it is also a moral requirement for myself, which naturally has moral authority. Furthermore, B can be reduced to the conjunction of the following three sentences: C = "Others are in our eyes reluctant to be the cause of the situation we do not want to see;" D = "In the eyes of us and others, we are loath to be the cause of conditions we do not want to see," and E = "In our opinion, we share with others the causal explanation after the first 'in our eyes' in D." Besides, E can be reduced to "$F \wedge G \wedge H \wedge I$...." In a sense, we can equate this nesting situation with the infinite expansion of *reflexive transitive closures*.

Assuming that public knowledge is a closed first-person *iterative combination* (which means a complete second-person standpoint), the amount of information in the *reflexive transitive closure* of a relationship tends to be infinite in the absence of any public knowledge (suppose moral knowledge is a networked structure). This problem is similar to the *mud baby puzzle*: if the father does not give any public knowledge, the two babies will never find out whether they have mud on their foreheads. There are two schemes to solve this problem: (1) Abandon the self-knowledge representation framework and interpret the above nesting situation as synonymous repetition; (2) Abandon completeness and consider the breaking of the cycle as the subject's incompleteness belief. Although scheme 1 retains the second-person cycle, its knowledge representation framework is third-person. Scheme 2 retains the second person to a certain extent, but breaks the cycle and also gives up completeness.

However, as experience evolves, normative content can evolve under an open first-person knowledge framework. If taking the scheme 1, we will face two issues. First, distributed morality will lose its significance because the tension between different definitions no longer exists. Second, Darwall's practice of classifying some abnormal normative behaviors as third-person perspectives will no longer be of any ethical or distinguishable representational significance. The "Self-conceit" phenomenon specifically labeled by him refers to the phenomenon of equating self-perspective norms with objective norms. In his view, his own will is the source of norms, namely the premise of egoistic ethical egoism (Darwall 2006, 136). Therefore, we recommend scheme 2, signifying that we must use a certain Defeasible Inheritance network with a non-monotonic reasoning mechanism. The requirements of system 1 are as follows: The inheritance network is a directed acyclic graph; the uncertainty of the inheritance relationship can be determined by some kind of "inherited truth function." Assuming that the inherited truth function is f(x), the conceptual inheritance diagram of the dual agent is shown in Figure 4 (the figure merely presents one truth function).

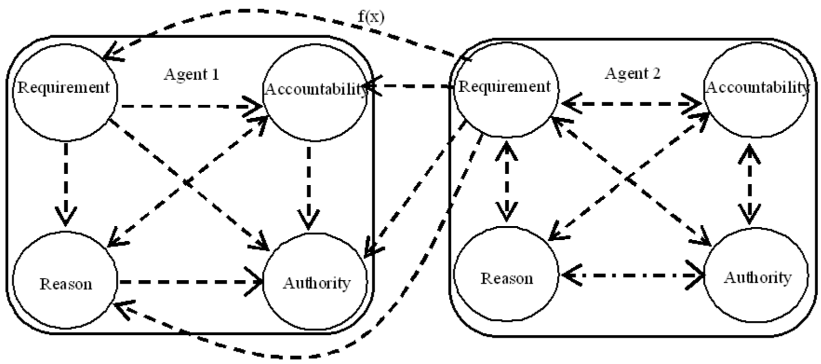

Fig. 4. Schematic diagram of the conceptual network between two agents, with the dashed lines indicating incomplete inheritance

3. Psychological distance and construal level theory

After establishing a fundamental conceptual inheritance relationship framework between subjects, the next step is to establish empirical quantification of the characterization for *inheriting functions*. This system could be set up in many ways, depending on different cognitive science achievements. However, considering the issue of adaptability, it is currently suggested to place this function under the framework of the mindreading theory. According to Bello and Bringsjord's theory (2012), disparate psychological distances can be analyzed qualitatively or quantitatively through the interpretation of construal level theory (CLT). Based on Knobe and Nichols's measurement of "causality involving self" (Knobe & Nichols 2011), a correlation analysis between different forms of causality is established in line with different types of self and the richness of cognitive content. Bello and Bringsjord suggested using CLT for reinterpretation in the light of their data, directly deeming different types of causality as different levels of abstraction. Hence, a correlation between the length of the causal chain and data that represent selves is established.

Originally proposed by Yaacov Trope and Nira Liberman (2010), CLT is a widely proven theory, asserting that similar people will have dissimilar cognitions of an identical event under different *psychological distances*, which is reflected in different construal levels of the event. The recognition of events with a longer psychological distance is more abstract and essential (which means a high construal level); the recognition of events with a shorter psychological distance is more specific and direct (which means a low construal level). The characteristic of psychological distance is not fixed, although mainstream items are time (close or not), space (close or not), the closeness of social relationships, and real or counterfactual events. Under the right conditions, the relationship between psychological distance and the construal level is symmetric. In other words, we can infer the construal level through psychological distance, and vice versa. Moreover, psychological distance is the "distance" between an event and the "self." Since the "self" is an abstract and complex concept, the "distance" from it may have many dimensions. Knobe and Nichols divide the concept of the "self" into three categories: the physical concept of the self, the psychological concept of the self, and the executive concept of the self. The psychological concept of the self refers to the mental process in the sense of psychology or cognitive science, which is the kind of "mental events" that philosopher Donald Davidson (2001) or Jaegwon Kim (1973) worked on. The corresponding causal relationship is *mental causation*. The concept of self-implementation refers to an "agent," a similar "subject" commonly used in the history of philosophy, and its metaphysical properties are not limited by physical time and causality (in contrast to from the psychological concept of the self). The corresponding causal relationship is *agent causation* (Clarke 1993; O'Connor 1996).

To find out which causal relationship is more easily used by people at different psychological distances, Knobe and Nichols tried to analyze target events through questionnaires. Once causation involving the "self" is associated with psychological distance, we can use this as a basis to establish a psychological distance between individuals. When the target event happens to someone else, the causal judgment we make is about the relationship between "the self of the other person and the event." Then, causal judgments about ourselves and different others can be made—listed as an extension of inherited functions. The detail is as follows: When zooming out and examining a person's behavior in a broader context, people are apt to view emotions or the body as part of the self. At this time, people are not explicit about the concept of the self, and are inclined to use the concept of the "executive self" to explain behavior in an abstract and unified way. When zooming in and examining a person's behavior in a more specific situation, people are likely to distinguish the concept of the self in a relatively clear way, such as not considering emotion as the "self" part. When the psychological process is separated from the "executing self," a psychological process rather than the agent (a "thin" self) may be regarded as the cause of the behavior (Knobe and Nichols 2011). Strictly speaking, the *uncertainty* concerning the causal attribution of the self regarding others' behavior can be perceived as the *extension* of the inheritance function.

Nevertheless, Bello and Bringsjord have another plan. They use the length of the causal chain to represent different psychological distances and use the level of abstraction to explain why humans have different causality under various psychological distances. This plan may lead to the immaturity of the mindreading representation. According to CLT, a higher level of interpretation means more abstract, vague, and inaccurate features, which is an economic thinking principle. A series of actions are taken as: "I want to eat sandwiches, I intend to make them. After opening the refrigerator, I take out various materials to make sandwiches, and later a sandwich appears on the table." When attributing this behavior, I tend to use the distal "I" as the reason for "a sandwich appeared on the table," for we don't have time to cope with so many reasons. In this way, the distal cause can easily be wrapped in an "agent." However, in cases where the psychological distance is shorter, the subject has sufficient resources to analyze the proximal situation, which explains the lower construal level of the relationship. For example, more people will tend to think that the reason for "the refrigerator is opened" is "I want to eat sandwiches" rather than "I" when the previous case is changed into "As I want to eat sandwiches, I intend to make them, and the refrigerator is opened."

Please note that the so-called "I" here does not mean the "first-person" as defined in section 2 but a specific instance of knowledge representation. From a philosophical point of view, this is a configuration of knowledge content based on "de se." There are many theories on this issue, and Bello and Bringsjord may pay more attention to Neil Feit's theory (Feit 2008). Feit tried to defend David Lewis's content-property theory, claiming that all psychological content is property, not a proposition. That means that all material of this statement is merely supervenient on the actual physical cause and effect. In "my first-person standpoint," the psychological content of others is characterized by counterfactual statements of my psychological content. When considering what others think, we often "simulate" others' thoughts as "if I were him, I would do Φ" on which the assumptions of "I" for others' thoughts are built. My counterfactual cognitive content will be different form "his" in "my first-person standpoint." The difference can be precisely described by psychological distance. In this sense, some hypothetical functions can be recursively produced based on research data like that of Knobe and Nichols to describe the quantitative relationship between psychological distance (causal chain length) and different simulation content. Furthermore, many specialized learning algorithms can be applied here. In summary, if the above model is used to construct intersubjectivity, the function used to describe this property is based on the environment. In other words, it has no direct relationship with the subject's internal history or various intricate internal cognitive properties.

4. The knowledge representation of mindreading under the framework of Polyscheme

The artificial intelligence framework used by Bello and Bringsjord is derived from the Polyscheme system built by Nicholas L. Cassimatis, who strove to build an "integrated" AGI system, referring to the integration of different specialist subsystems into a unified system with general capabilities. Given that different specialists can normally deal with a particular type of work or one or a few aspects of an event, we can understand an aspect as only dealing with a specific property of an event. The purpose of Polyscheme is to combine different property processing capabilities so that the machine can handle complex and multifaceted events like a human—a comprehensive property processor. The system presents a "foci" state, which is a proposition that can be used to connect all specialists. For the sake of the synergy between different specialists, Cassimatis set up many protocols, thereby ensuring that "all specialists focus on the same proposition at the same time." In line with a "de se" representation, Polyscheme may presuppose incomplete semantics, signifying that some contents in this world are propositions or can at least be represented by propositions. Besides, it also means an objective representation. As we will see later, this allows the self-knowledge representation that is constructed in the Polyscheme system to have some properties that do not meet our expectations in the second-person analysis.

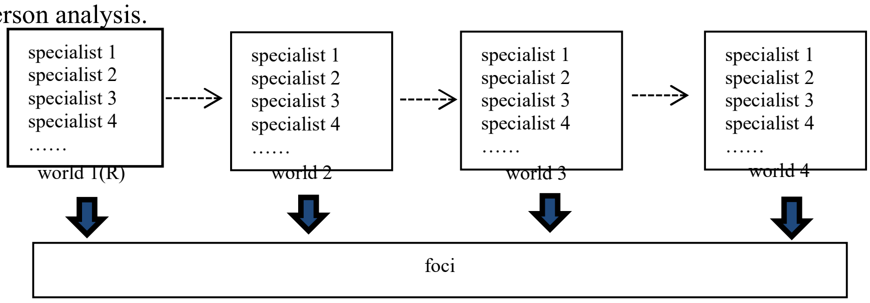

Fig. 5. The dashed line indicates the inheritance relationship between possible worlds, with the down arrow pointing to the proposition

The foci proposition is the kernel of the specialist net and includes four elements: a predicate, zero to multiple arguments about the predicate, the interval tense, and possible-world semantics. The last element can be used to settle the entire range of counterfactual problems, uncertainty problems, and "other mind" problems called the "mindreading" problem (Cassimatis 2001). The R world is the real world defined by Polyscheme, and other worlds are a subset or a superset of R. What is the target world "?W" can be characterized as more false or real (the relevant framework is shown in Figure 5). When all the ?W is inherited to R, there are two sorts of inheritance relationships. One is hard inheritance, with the symbol "==>," analogous to modal necessity. Suppose E designates the temporal interval representing "at all times," then the reasoning "If John is happy, he is Full." is written as "Happy (john, E, R) ==> Full (john, E, R)." The other is soft inheritance, with the symbol expressed as "(cost)>." The value of cost ranges from 0 to 1. Generally, it is expressed as: Happy (john, E, R) (.80)> Full (john, E, R), and the justification of this reasoning is a problem of degree (80 is a default value). Weak inheritance is employed to help crudely describe the psychological distance and causal chain length related to mindreading. Treating different agents as disparate worlds, we then get the following representation of the inheritance relationship concerning relational predicates in the eyes of different subjects at any moment (Let ?Relation be any relationship, here it can be causal relationship; ?e1 is any entity, and t1 is any time):

?Relation(e1;?t; R)∧IsCounterfactualWorld(?w, E, R) (cost) >?Relation (e1;?t;?w) (Bello and Bringsjord 2012).

The setting of cost is: the larger the value, the closer the counterfactual world of the mindreading object is to the real world (that means from the first-person perspective). The value is related to the length of the causal chain. The specific constraint judgment threshold is given by the external CLT data. Although this is not sufficient to serve for the composition and reasoning of the second-person distributed knowledge structure we mentioned, it may be hoped to build a basis for it. At least, the aforementioned inheritance relationship can assist the subject in recognizing the distance between the moral reason of others and his own. Hence, the similarities and differences between his own responsibility view and that of mindreading can be evaluated.

5. The moral machine under the Polyscheme framework from the perspective of the construction concerning intelligent agents

This moral machine under the Polyscheme framework currently appears to have two features: (1) It is an integrated intelligence agent; (2) its inter-subject inheritance function stems from the induction of external data. It will be found that the first feature needs the second one to cooperate with it to build this moral machine. Such a configuration will face some essential questions from the level of artificial intelligence and philosophy.

As we see, the mindreading representation of Polyscheme is counterfactual reasoning, which involves relationships between different worlds. The worlds here refer to the agreements reached by all specialists concerning representations of the same foci proposition. Cassimatis's construction of cognitive subjects is based on Baars's theory (Baars 1993). He believes the human mind has many different specialized systems, which generally exist subconsciously and do not need to communicate with each other. This provides theoretical space for the "integration" of specialists with different representations. According to Baars, consciousness is a kind of "global workspace" that provides a channel for information exchange between different specialized systems. Being more flexible than the AGI model of LIDA built with GWT (global workspace theory) as a template (Baars and Franklin 2009), the foci proposition is a similar global workspace that directly coordinates the specialists on the foci events.

In general terms, global workspace is more like a conference room. Different specialists send and receive information to determine to which part their "attention" should be paid in the proposition. For example, (Cassimatis 2001) there is a system that first recognizes Joe and then Red. What needs to be determined is whether they are the same person. Suppose the system has three specialists performing tasks: a physical location recognition system (a), a color recognition system (b), and a sounding system (c). At this point, (a) and (b) will contradict each other. According to (b), it can be concluded that Red moves from the opposite bank of the river to Joe across the middle but finds that the color of Red (covered in mud) differs from that of Joe. At this time, the higher-level cognition of the system is in need of judging whether they are the same person, in line with the information given by (a), (b), and (c). By this time, (c) is a core system, and the information it gives helps determine that they are an identical person. However, assuming that the colors on Red and Joe are the same, (c) will ask (a) and (b) to review, and even ask (a) to expand the search to the entire map to see if someone has replaced Joe across the bridge. In this case, (b) becomes a core system again. This is simply an example of visual reasoning, and it will involve the highly complex coordination problem of all attribute processors in the framework of mindreading. In reality, different "worlds" mean collaborative contracts for different dedicated systems, and various contracts signify disparate cognitive conditions. To deal with non-coordination, there is a reflective specialist system specifically used to find inconsistencies between the worlds and the resolution of these inconsistencies, thereby forming a "meta consensus" (Cassimatis 2001). It can be seen from the mindreading application that the inheritance function of the external psychological distance can be used as one of the criteria for planning the meta-consensus.

According to the operating mechanism of meta-consensus, Polyscheme seems to be able to monitor external meta-consensus parameters, but it is indeed difficult to complete. The first reason is that Polyscheme's supervision method rests upon frequency supervision to complete the supervision of inherited function reliability. This makes it probable to detect the system's success in obtaining the expected feedback based on the consensus, but the fitting of the inheritance function (assuming that an object is a real person) is a highly statistical result that is based on the actual situation. The convergence of the function is reliable under the same environment, and thus it may be hard to find inconsistencies. The second reason is that under the premise of subject interaction, supervision is sometimes realized through interpersonal interaction and feedback.

Nevertheless, the interaction itself may change the mind of the object, thus breaking the intersubjectivity that has been established before. We can call this interference "*interactive interference*." Moreover, ordinary people do not rely entirely on external parameters to guide their cognition. In fact, humans often cannot directly acquire inter-subjective parameters that require complex algorithms to be obtained from the outside world. Instead, some internal overall cognitive experience is often employed to judge the relationship between subjects. We can call this the structure of the subject's cross-temporal identity, or some kind of "personality" or "character," which will be accompanied by the subject's unique way of interacting with the outside world and its judgment criteria. Sometimes human beings judge other people's behavior and inner thoughts through "subjective" methods; this subjectivity is not random but condenses cognitive history development. In other words, the judgment mechanism of meta-consensus should not just originate from the outside world.

However, the formation of meta-consensus is compatible with Bello's computational account of mindreading. The mindreading representation is ascribed to the third-person stimulation, which seems to not match the requirement of the *second-person concept cycle-distributed network*. The third-person stimulation means treating the cost of the inherited function as an objective criterion for the counterfactual inference of mindreading, so the system can "introspect" about the other minds to an exact degree. Nevertheless, mindreading is not necessarily a third-person cognitive process, and so is the self-knowledge about future or past. There is no such universal knowledge unless we presuppose it. And we cannot accept it based on the argument in section 2.

Due to the principle of the knowledge representation and control of Polyscheme, it may have two drawbacks. Its first flaw is to build the self with the event rather than with "history," which means that the self is fragmented and the continuity between egos cannot be guaranteed. A "self" that is not lost in the experience should have a unified representation system and an internal operation mechanism. The experience content displayed by this mechanism can change with the variation of external input experience, but the core mechanics will be invariable. In a sense, this may be some kind of "transcendent" self. According to Piaget's theory (Piaget and Cook 1952), the process of children's mental growth is gradually completed by combining specific perceptions and behaviors with the reflective operation of the original self-mind. Thus, merely a continuous self can guarantee the continuity of moral judgment.

Another drawback of Polyscheme is that its *set of basic possible-world semantics* cannot describe all the distinction and closeness of all possible worlds. And the missing ones may mean those implied by some realistic psychological distance. Here, we introduce David Lewis's possible-world theory to illustrate how the missing ones regularize. According to Lewis's theory, qualitative similarity is more vital than quantitative similarity in theory. Lewis has an analysis of counterfactual dependency causality: "If A, then C." is true if and only if the world in which A and C are both true is more similar to our real world than the one in which A is true but C is false (Lewis 1979, 465). We can write the world in which A and C are both true formally as "W(A&C)," the world in which A is true but C is false as "W(A&¬C)," and the real world as "W(¬A&¬C)." Then, we have $W(A\&C) <_{R(\neg A\&\neg C)} W(A\&\neg C)$, which can be transformed into $W(A\&\neg C) <_{R(\neg A\&C)} W(A\&C)$. From a quantitative perspective, W(A&C) is farther from the real-world R(¬A&¬C) compared with W(A&¬C), but closer to it in a qualitative sense. Let's use an example of a topology diagram to explain the meaning here (see Figures 6–8). Although Figure 7 resembles Figure 6 more than Figure 8 from a quantitative perspective, Figure 8 is more similar to Figure 6 in comparison with Figure 7 from a qualitative angle. Thus, we have Figure 8 $<_{\text{Figure 6}}$ Figure 7.

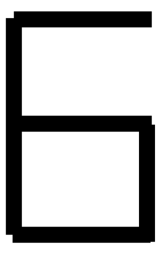

Fig. 6.

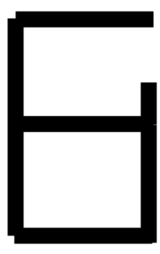

Fig. 7.

Fig. 8.

World similarity in Polyscheme is built on the basis of sets. And in the knowledge representation of mindreading, similarity is constructed on the scale of psychological distance (causal chain length). Suppose that the *increasing* of psychological distance (the lengthening of the cause chain) is set to A, and the *recognition* of some abstract agent causation is set to C (there is no trend change in "real-world" R. For the convenience of processing, we define the decline of the trend as disapproval and the increase as approval). The experimental elements and final results in Knobe and Nichols's paper (2011) are as follows:

	Background (assuming you observe the following scenario).	Questions (scores 1–7).	Average scores of answers
Choice-Cause (Zoomed-In) (1)	A bee landed next to John, whose hand is drawn back. You know John pulled back because he was afraid of bees.	John caused his hand to withdraw.	6.10
Choice-Cause (Zoomed-Out) (2)	A bee landed next to John, whose hand was drawn back, and a bottle of milk was knocked over. You know John pulled back because he was afraid of bees.	John caused the milk to spill.	4.95
Emotion-Caus **(Zoomed-In) (3)**	A bee landed next to John, with his hand shaking. You know John's hand shakes because he's afraid of bees.	John caused his hands to shake.	3.95
Emotion-Cause (Zoomed-Out) (4)	A bee landed next to John, with his hands shaking, and a bottle of milk was knocked over. You know John's hands shake because he's afraid of bees.	John caused the milk to spill.	5.48

Table 1.

In the relationship between two items of selective attribution ((1)→(2)), there is A&¬C. Besides, the relationship between the two items of emotional attribution ((3)→(4)) is A&C. We set the former as Wc or W(A&¬C) and the latter as We or W (A&C). Suppose that the *basic* first-person mental state is R, and the variation of the psychological distance in R is zero. According to the knowledge frame of mindreading, if the self is added to the reasoning, there will be C. Therefore, we tend to think that there is ¬A&C in R. Based on Lewis's theory, as long as W(A&¬C) $<_{R(\neg A\&C)}$W(A&C) is true, we find a counterexample of the mindreading representation concerning Polyscheme. The question now is: Is Wc more similar to R than We? We are inclined to think so.

First, according to (1)→(2), we can see that the lengthening of the causal chain does not necessarily mean the change in psychological distance predicted by C, which shows that the world mechanisms of Wc and We may be different. Second, in terms of selecting attribution, people tend to choose “abstract attribution” from the beginning when there is extremely close psychological distance, which is very close to the state of complete rationalization in R. Even if the cause-effect chain grows, we believe that people will still be apt to perform causal attribution on the target object, and the average answer declines not because people have changed the subject attribution method. Instead, they have joined the investigation of the contingency, and the subject’s choice obviously does not cover the milk splash at the far end, which may be defined by people as an accident. This attribution mechanism is clearly more similar to that in the R world. Therefore, we believe that Bello and Bringsjord have simplified the similarity problem in the possible world, and the mindreading representation system has overlooked some experimental results and the significance of Knobe and Nichols’s original analysis.

6. Self and mindreading

If the process of internal consultation in Polyscheme’s systems is compared to a meeting, it means that the power to finalize the decision comes from a person in a high position, although his/her decision-making reference comes from some external force. From this point of view, the control of reflective specialists is too strong. Even if the external parameters are relatively objective, the framework setting of the internal possible worlds also results in the situation that parameters are not conducive to establishing a comparatively complete mindreading representation. In addition, it is impossible to establish a complete representation of mindreading knowledge in a sense—at least the *interactive interference* mentioned above may be difficult to avoid. Some people may think that we can try to build a complete system of possible worlds, and then use it as a basis to fit a complete mindreading function. However, this is also a “top-down” approach, which often encounters various types of frame problems in line with judgments by Wendel Wallach, Colin Allen, and Franklin (2010).

The core of the problem may not be to establish a complete mindreading frame but to establish a representation that can give timely feedback and adjust the model. Moreover, this is related to the self-knowledge representation of the AGI system and the system’s setup ideas as well. A way of avoiding Polycheme’s problems can consist of the following: (1) Establish a unified inherent kernel; (2) set up an adaptation mechanism concerning the environment. This requires the system to have an internal control subsystem that can characterize the tendencies within the system and vary as the internal experience changes. Apart from that, the system does not need to pursue too much consistency of internal knowledge, like Polyscheme (although the mechanism of knowledge representation is unified, see Figure 9), and thus it is not essential to always give a “self” with coordinated functions. The internal control subsystem and external feedback can determine a “self” state when necessary.

In this sense, it fits the vision of Figure 3. Besides, we can design a mindreading representation, as in Figure 10.

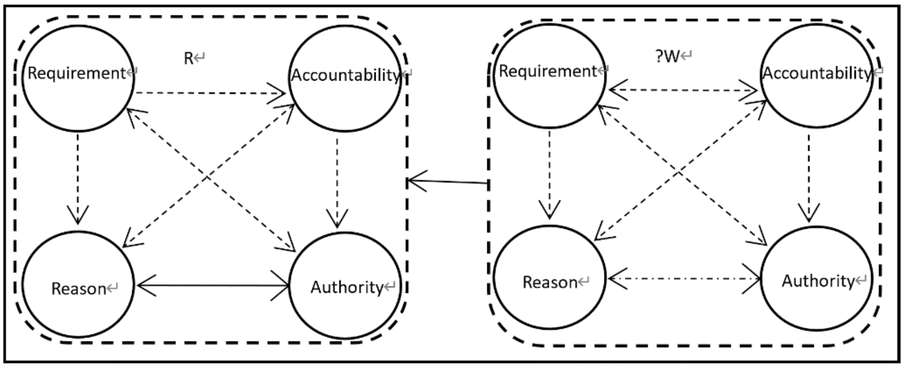

Fig. 9. The dashed boxes represent system boundaries, and the dashed boxes represent different sub-world boundaries

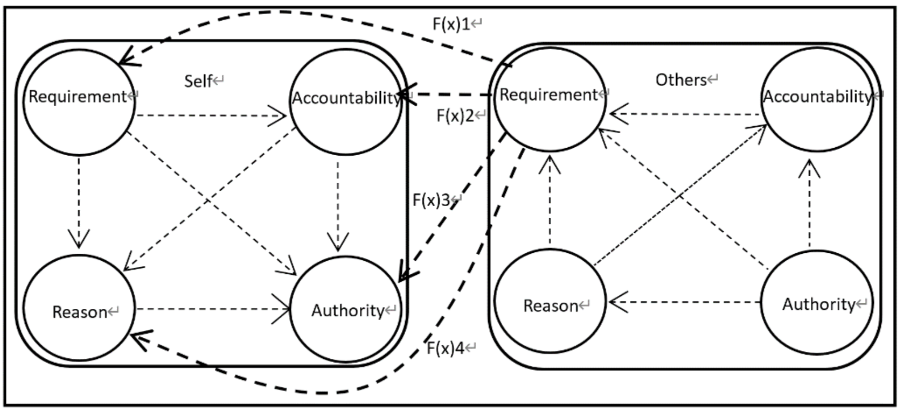

Fig. 10.

There are few systems in the AGI field that meet the above requirements, and Wang Pei's NARS (the Non-Axiomatic Reasoning System) is considered a potential candidate (Wang 2006). The system follows a kind of non-public logic to establish inference rules, and its characterization of "propositions" is accomplished through the inheritance relationship between terms. The core of NARS is not a multi-tasking event processor composed of specialists, but a belief network made up of many concepts, whose connotations and extensions are continually changing. NARS follows experience-grounded semantics (EGS), the basic premise of which is "*relatively insufficient knowledge and system resources*" (AIKR) that prevent NARS systems from encountering framework problems. The choice of experience will be based on the system's preference settings—motivation and emotion. It can also determine the representation of self-knowledge, and then the representation of mindreading. Thanks to AIKR, NARS does not need to maintain a constant consistency internally. With the input of external experience, the concept network will change, and unsuitable beliefs will be deleted in principle. Therefore, the relationship between beliefs does not need to be an intact inheritance relationship. Whether it is external information or internal information, it can affect the target concept or belief through intension or extension in the net on the whole. Thus, there may be no need to presuppose the semantic structure of the possible world in NARS to deal with mindreading problems, because there is no explicit conceptual boundary between the self and others.

7. Conclusions

Through our elaboration and analysis, we can draw at least the following conclusions: (1) Darwall's second-person moral philosophy leaves much space for the distributed expression of intersubjectivity and second-person knowledge. (2) Using a single mindreading technique or theory to formulate an inheritance function and using it as an "objective" reference standard to establish a mindreading knowledge representation may run counter to the assumption of second-person moral philosophy. (3) Since this type of AGI system concerning Polyscheme will face various issues in achieving the ideal mindreading representation, an AGI system with unified internal knowledge representation that can avoid the framework problem may have more potential to be applied to moral machines to realize the vision of Darwall.

References

1. Asady, R.: Inheritance theory: an artificial intelligence approach. Intellect Books, Bristol (1995).
2. Anderson. M., Anderson, SL., Armen, C.: MedEthEx: a prototype medical ethics advisor. In: Proceedings of the National Conference on Artificial Intelligence, Menlo Park, CA(2006).
3. Baars, BJ.: A cognitive theory of consciousness. Cambridge University Press, Cambridge(1993).
4. Baars, BJ., Franklin, S.: Consciousness is computational: The LIDA model of global workspace theory. International Journal of Machine Consciousness 1(01):23–32 (2009).
5. Bello P, Bringsjord, S. On how to build a moral machine. Topoi 32(2):251–266 (2013).
6. Cassimatis, NL.: Polyscheme: A cognitive architecture for intergrating multiple representation and inference schemes. Dissertation, Massachusetts Institute of Technology, Cambridge(2001).
7. Clarke R Toward. a credible agent-causal account of free will. Noûs 27(2):191–203(1993).

8. Dameski, A.: A comprehensive ethical framework for AI entities: Foundations. Paper presented at the International Conference on Artificial General Intelligence, Prague, Czech Republic, August 22–25 (2018).
9. Darwall, SL.: The second-person standpoint: Morality, respect, and accountability. Harvard University Press, Cambridge(2006).
10. Davidson, D.: Essays on actions and events: Philosophical essays volume 1: Philosophical essays. Oxford University Press on Demand, Oxford(2001).
11. Feit, N.: Belief about the self: A defense of the property theory of content. Oxford University Press, Oxford(2008).
12. Gilligan, C.: In a different voice: Psychological theory and women's development. Harvard University Press, Cambridge(1993).
13. Guarini, M.: Computational neural modeling and the philosophy of ethics reflections on the particularism-generalism debate. In: Anderson M and Anderson S (eds) Machine ethics 316–334. Cambridge University Press, Cambridge(2011).
14. Kim. J.: Causation, nomic subsumption, and the concept of event. The Journal of Philosophy 70(8):217–236(1973).
15. Knobe. J., Nichols. S.: Free will and the bounds of the self. In: R Kane (ed) The Oxford handbook of the free will: Second edition. Available on PhilArchive: https://philarchive.org/archive/KNOFWA(2011)
16. Lewis, D.: Counterfactual dependence and time's arrow. Noûs 13(4):455–476(1979).
17. McLaren, BM., Ashley. KD.: Case-based comparative evaluation in TRUTH-TELLER. In: Proceedings From the Seventeenth Annual Conference of the Cognitive Science Society, Pittburgh, PA(1995).
18. O'Connor ,T.: Why agent causation? Philosophical Topics 24(2):143–158(1996).
19. Piaget, J., Cook, MT.: The origins of intelligence in children. International Universities Press, New York(1952).
20. Rawls, J.: Kantian constructivism in moral theory. The journal of philosophy 77(9):515–572(1980).
21. Savulescu, J., Maslen, H.: Moral enhancement and artificial intelligence: Moral AI? In: Romportl J, Zackova E, Kelemen J (eds) Beyond artificial intelligence. Topics in intelligent engineering and informatics, vol 9. Springer, Cham (2015).
22. Strawson PF Freedom and resentment and other essays. Routledge, London(2014).
23. Trope, Y., Liberman, N.: Construal-level theory of psychological distance. Psychological review 117(2):440–463(2010).
24. Wallace, RJ.: The moral nexus (Vol. 9). Princeton University Press, Princeton(2019).
25. Wallach W, Allen, C.: Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford(2008).
26. Wallach, W., Franklin S, Allen C A conceptual and computational model of moral decision making in human and artificial agents. Topics in cognitive science 2(3):454-485(2010).
27. Wang, P.: Rigid flexibility: the logic of intelligence (Applied Logic Series 34). Springer Science & Business Media, New York (2006).
28. Williams, B.: Shame and necessity (Vol. 57). University of California Press, Berkeley(2008).