Logic of Knowledge and Cognitive Ability

Jia Tao
Lafayette College
Easton, PA, United States
taoj@lafayette.edu

Xinran Zhang
University of Illinois Urbana-Champaign
Champaign, IL, United States
ranxinzh@gmail.com

ABSTRACT

Along with the convenience brought by the increasing usage of autonomous systems, unexpected accidents happened. These accidents emphasize the need for autonomous systems to possess the ability to recognize potential hazards, effectively communicate these hazards to human operators, and facilitate retrospective analyses. This ability, including recognition, prejudgment, post-analysis, and reasoning, falls within the realm of cognitive ability, which is important in improving the safety and outcomes in the decision-making process of such systems. In this paper, we present a foundational step toward addressing the safety challenges involving the cognitive ability of artificial agents. We study the interplay between knowledge and the cognitive ability of intelligent agents. The main technical result is a sound and complete bimodal logical system that describes the interplay between the knowledge and cognitive ability modalities.

KEYWORDS

epistemic logic, cognitive ability, axiomatization, completeness

1 INTRODUCTION

Self-driving cars are gradually becoming part of people's everyday lives. However, safety concerns are rising along with the technology advancements. Designing self-driving systems equipped with the ability to provide safe and reliable driving experiences is a complicated task. A federal government's top auto-safety regulator disclosed in 2022 that "nearly 400 crashes in the United States in 10 months involved cars using advanced driver-assistance technologies" [6]. Past accidents show that some autonomous systems are still questioned to be flawed. For example, Autopilot, Tesla's driver-assistance system, was involved in "three fatal crashes occurred in a 51-day span" in the summer of 2022 with a similar pattern: "a person driving a Tesla in the early morning hours with Autopilot active strikes a motorcycle" [29]. These crashes raise concerns among motorcycle advocates. They worry that Autopilot may be incompetent in fully recognizing motorcycles while lulling its drivers into a false "sense of complacency and inattentiveness" [29].

This false sense of security in drivers, believing their autonomous cars drive by themselves, also led to other accidents unrelated to motorcycles. For instance, in 2020, Rafaela Vasquez, the backup driver of an Uber test vehicle operating in autonomous mode, was charged with negligent homicide after the Uber vehicle struck and killed a pedestrian in Arizona in 2018 [28]. In 2023, in San Francisco, a Tesla Model Y, allegedly in self-driving mode, failed to slow down while a school bus displayed its stop sign. Consequently, a teenager was struck and thrown into the windshield, flew into the air, and landed in the middle of the road [41].

These incidents highlight significant issues in the current capabilities of self-driving technology. Autonomous cars are usually equipped with advanced technologies to assist in driving and avoid collisions. These advanced technologies provide self-driving systems with cognitive abilities to acquire information (i.e., the detected objects or other cars' speed and inertia), reason about the obtained information, and make decisions. Unfortunately, accidents still happen. How can self-driving systems be designed to better prevent accidents and reduce crashes? Let us look into these accidents more closely.

The 2018 Uber accident accident occurred when the victim, 49-year-old Elaine Herzberg, was slowly walking with a bicycle to cross a major road. The autonomous car did not slow down or change its course to avoid the pedestrian, resulting in a fatal crash [39]. It was reported [5] that the self-driving system indeed detected the pedestrian with her bicycle six seconds before the crash; however, it classified her as an unknown object. Unable to correctly recognize the pedestrian and her bicycle, the self-driving system did not take any action to actively avoid the pedestrian. When the system finally determined that emergency braking was required, it was only 1.3 seconds before the impact, too late to avoid the crash. From the report [28], the self-driving system relied on the driver to intervene and take action in emergencies, but it was not designed to alert the driver. The driver, Rafaela Vasquez, without receiving any warning, was watching television on her smartphone when the pedestrian was struck by the car, directly causing the pedestrian's death.

In this accident, Rafaela Vasquez was charged with negligent homicide. However, if the system had created a warning (e.g., a sound with a message on the display screen) right after it detected the pedestrian, the driver might have had enough time to react to prevent the tragedy. Even when a self-driving system is uncertain about an undesirable outcome, it should still warn the driver regarding any detected "unknown" information instead of assuming it had permission to proceed, particularly in situations where driver intervention may be required. Just as Floridi and Sanders [12] argued that one of the guidelines for a moral intelligent agent is interactivity – the ability to respond to stimulus, *a reliable self-driving system should have the ability to provide warnings when foreseeing a potentially undesirable outcome.*

For self-driving systems to have such an ability, they must not only be aware of the tasks they are performing, such as accelerating or slowing down, but also of the contextual conditions that dictate how these tasks should be executed. For instance, a self-driving car must know that it needs to stop when approaching a school bus displaying a stop sign, but not when approaching a temporarily parked Amazon truck. Thus, it is critical that *the system should know its own capabilities.* Let χ represent the statement "hitting an obstacle", the modal formula $\mathsf{A}_{auto}\chi$ denote the statement that "the autonomous car has the ability to foresee that the current action leads to the outcome χ," and K_{auto} denote "the autonomous car knows that". Then, the following formula holds:

$$\mathsf{A}_{auto}\chi \rightarrow \mathsf{K}_{auto}\mathsf{A}_{auto}\chi \tag{1}$$

Once the system concludes $\mathsf{A}_{auto}\chi$, it knows $\mathsf{A}_{auto}\chi$ and should alert the driver about χ. Note that such communication should occur before the potential accident to serve as a warning. Thus, the knowledge modality K is the *ex ante* knowledge, i.e., the agent's knowledge before the event happens.

Another important feature that self-driving systems should have is *the ability to provide information for retrospective analysis.* In the 2023 Tesla accident, it was unclear why the self-driving Tesla never slowed down while the school bus displayed its stop sign. Similarly, the recurring pattern of Autopilot hitting motorcycles also lacked a clear explanation. As autonomous agents are rational, they are designed to take the best possible action given any situation. However, if the "best" possible action results in an accident, it often means the system missed certain critical information and made a wrong decision, leading to an unfavorable outcome. To prevent similar outcomes in the future, it is critical for a self-driving system to have the ability to provide information for an undesirable outcome. This retrospective analysis can help designers identify and rectify the underlying issues, thereby enhancing the safety and reliability of the system.

Let φ represent "pedestrian is hit by a car", ψ represent "car keeps driving" and modality A represent "have the ability to provide information/explain/conclude". Normally, when a pedestrian is crossing the road, the autonomous car knows that if the car keeps driving, it will hit the pedestrian. This is represented by $\mathsf{K}_{auto}(\psi \rightarrow \varphi)$. Yet, in the 2023 Tesla case, despite knowing that driving (ψ) would lead to hitting a pedestrian (φ), the car kept driving. This implies that the Tesla system considered the situation safe. Thus, if the autonomous car has the ability to provide an explanation for its actions ($\mathsf{A}_{auto}\psi$), then it must also be able to provide an explanation for the consequences of those actions ($\mathsf{A}_{auto}\varphi$). The following logical expression formalizes this relationship:

$$\mathsf{K}_{auto}(\psi \rightarrow \varphi) \rightarrow (\mathsf{A}_{auto}\psi \rightarrow \mathsf{A}_{auto}\varphi) \tag{2}$$

According to Tesla's manual [27], drivers have the option to enable a feature in the Autopilot of Model Y, the same model involved in this accident, which allows the car to automatically stop for red lights or stop signs. It remains unclear whether this feature was enabled in the vehicle at the time of the incident. If it was not enabled, the system should be designed with the ability of providing warnings to the drivers, enhancing the safety and reliability.

On the other hand, consider a situation when the feature for detecting red lights or stop signs is enabled. If the Autopilot does not recognize any red light or stop sign, then the car will keep driving. Let ψ_1 denote "no red light or stop sign is detected" and ψ_2 denote "the feature is on". Recall that ψ represents "car keeps driving". Normally, cars can keep driving when there is no red light or a stop sign detected while the feature is on. If the autonomous car has this knowledge, then its ability to provide an explanation for why no red light or stop sign is detected while the feature is on implies its ability to provide an explanation for keeping driving. This can be expressed as:

$$\mathsf{K}_{auto}(\psi_1 \wedge \psi_2 \rightarrow \psi) \rightarrow (\mathsf{A}_{auto}(\psi_1 \wedge \psi_2) \rightarrow \mathsf{A}_{auto}\psi)$$

In general, it is reasonable to assume that an autonomous car has the knowledge that it can keep driving when there is no red light or stop sign is detected while the feature is on. Thus, if the Autopilot has the ability to provide an explanation for not detecting the stop sign of the school bus while the detection feature is on ($\mathsf{A}_{auto}(\psi_1 \wedge \psi_2)$), then it has the ability to provide an explanation for the car to keep driving ($\mathsf{A}_{auto}\psi$), which led to the teenager being struck. By analyzing the system's ability to explain its actions, developers can identify that a self-driving car may not be able to recognize a stop sign on a school bus.

To summarize, it is crucial for autonomous systems to possess the ability to recognize potential hazards, effectively communicate these hazards to the driver, and facilitate retrospective analyses to enhance safety and reliability. This ability–including recognition, prejudgment, post-analysis, and reasoning–falls under the domain of the cognitive ability. Note that autonomous agents are built on complex architectures, consisting of separate modules for perception, data management, legislative requirements, and more. For example, a comprehensive analysis of communication with the human operator would also need to account for human factors such as the operator's awareness, workload, and attention. Our focus, however, is to study the fundamental properties of the interplay between cognitive ability and knowledge. A detailed analysis of the practical aspects of cognitive ability is beyond the scope of this paper. Our technical contribution is a sound and complete logical system for reasoning about the interplay between cognitive ability and knowledge.

The rest of the article is organized as follows. In Section 2, we discuss the related work and how we position our work in the literature. In Section 3, we introduce the syntax and the semantics of our logical system. Section 4 provides a list of axioms of the system and the proofs of their soundness. The completeness proof is presented in Section 5. Finally, in Section 6 we conclude.

2 LITERATURE

The concept of ability has been a topic of longstanding philosophical debate. Van Inwagen [46] argued that ability is a power that connects an agent to an action. Reid [36] suggested that ability shares similarities with the traditional notion of active powers, which involve the will. Ryle [38] linked the concept of ability to the idea of knowing how to perform an action, while Stanley and Williamson [43] contended that knowing-how is essentially a form of knowing-that. Kasirzadeh and McGeer [19] and Mele [30] distinguished between two types of ability: specific and general. Specific ability refers to the capacity of an agent to perform an action when all necessary conditions are met, whereas general ability refers to the capacity to perform an action even when not all conditions are favorable. Robb [37] explored how to integrate intelligent powers—such as skills or talents—into the account of ability, noting that these powers are linked to practical intelligence.

Few researchers have explored cognitive ability. Hernández-Orallo and Dowe [15] identified key differences between the cognitive abilities of animals and machines: animals possess innate cognitive abilities, while machines' abilities are tied to interactive systems, with actions based on prior observations. Their study introduced the concept of potential cognitive ability, defined as "how quickly and likely the process of acquiring the ability is." They measured various machine characteristics and analyzed relationships between certain potential abilities, but did not address knowledge. Konek [20] took a quantitative approach, investigating probabilistic knowledge and cognitive abilities, and argued that cognitive ability involves reasoning about how evidence supports or undermines credences. Unlike both studies, our research explores the relationship between knowledge and cognitive ability, using a logical framework to formalize and reason about their interplay.

Pritchard [35] offered a philosophical discussion on the relationship between an agent's knowledge and cognitive ability, arguing that an agent can acquire knowledge through its cognitive abilities. Our study, in addition to being grounded in logic, differs from Pritchard's by focusing especially on the cognitive abilities of intelligent agents. In our approach, agents can not only acquire perceived knowledge as situations change but also possess predefined knowledge, such as road signs or maps.

Our definition of cognitive ability includes the capacity to provide information for retrospective analysis, which necessitates that intelligent agents possess the ability to offer explanations. The study of an intelligent system's ability to provide information or explanations for its actions dates back more than forty years [40, 44], to when expert systems were developed to explain why certain actions were taken. In [40], the ability to explain was achieved using production rules that encoded domain-specific judgmental knowledge, linking situations to actions. To explain the system's actions, it was required that the explanation system understand how or why certain rules were applied and maintain a comprehensive record of specific actions taken. This requirement is similar to our requirement that an intelligent agent should also know its own capabilities, as formalized in statement (1).

To formally define the concept of ability, some researchers have employed possible worlds/states semantics [9, 42]. A widely accepted view is that an agent possesses an ability if the agent performs the action in some possible world [42]. Another popular approach to the semantics of ability comes from a linguistic perspective [21–23], where possible worlds semantics is extended with contextual factors to analyze abilities within complete assertions.

Many logical systems have explored the interaction between ability and action. Alechina et al. [3] considered the situations where agents take actions under limited resources to achieve their goal. They extended coalition logic with resource bounds to describe both single-step strategies and multi-step strategies and gave a sound and complete axiomatization of the logic. Goranko [13] explored the intersection of coalition logic and alternating temporal logic (ATL) and provided a complete axiomatization. This work embedded coalition logic into ATL so that agents can coordinate strategies with time to achieve goals. Other works include STIT ("seeing to it that") logic, introduced by Horty and Belnap [18] and axiomatized first by Xu [47] and later by Balbiani et al. [4], as well as agency and deontic logic by Horty [17]. Temporal STIT logic was explored by Lorini [25], and the epistemic logic of blameworthiness was developed by Naumov and Tao [34]. Other contributions, such as those by Abarca and Broersen [1], Broersen [7, 8], focus on "seeing to it that" or responsibility-related concepts. However, none of these works considers the cognitive ability to provide information about actions that lead to specific outcomes, an essential aspect for enhancing the safety and reliability of intelligent agents.

Building on this line of research, we formally define cognitive ability as the ability of an agent, through its cognition, to perform an action leading to specific outcomes. Unlike previous work, we base cognitive ability on cognitive relations formed through information acquisition across possible states. Prior research [20, 35] has shown a strong connection between cognitive ability and knowledge, emphasizing the need to study their interaction for improving the safety and reliability of autonomous systems. To the best of our knowledge, we are the first to formalize a logical system that integrates both knowledge and cognitive ability, proposing a sound and complete system to reason about their interplay.

3 THE LANGUAGE

In this section, we introduce the formal syntax and semantics of our logical system. We assume a fixed set of propositional variables and a fixed set of agents $\mathcal{A}$. The language Φ of our logical system is defined by the grammar: for each $a \in \mathcal{A}$,

$$\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid \mathsf{K}_a\varphi \mid \mathsf{A}_a\varphi \;.$$

We assume that the constants $\top$ and $\bot$ as well as the Boolean connective $\wedge$ are defined in the standard way.

To interpret modality A, which represents the ability of an agent, under its cognition, to perform an action leading to an outcome, we define a set I of initial states, a set Δ of actions, a set Ω of outcomes, and a cognition relation $\sim_a$ for each agent $a \in \mathcal{A}$ in our model. To interpret modality K, following the epistemic logic S5, we use an equivalence relation $\equiv_a$ for each agent $a \in \mathcal{A}$. Recall that modality K represents the *ex ante* knowledge of an agent (see Section 1). Thus, relation $\equiv_a$ is applied on set I. We refer to our modal as a *game* because we focus on one-shot games. Since not all transitions from an initial state to an outcome via an action are valid, we define a set P of valid transitions, called *plays*. Additionally, the function π is used to interpret the propositional variables, mapping each propositional variable p into a set of plays where p is true.

For any set Δ of actions, we use $\Delta^{\mathcal{A}}$ to denote the set of all functions from set $\mathcal{A}$ to set Δ, which represents the set of all action profiles. The formal definition of a game is as follows.

Definition 1. *A tuple $(I, \equiv, \Delta, \Omega, \sim, P, \pi)$ is called a game, where*

1. *I is a nonempty set of "initial states",*
2. *$\equiv_a$ is an "indistinguishability" equivalence relation on the set I for each agent $a \in \mathcal{A}$,*
3. *Δ is a nonempty set of "actions",*
4. *Ω is a nonempty set of "outcomes",*
5. *$\sim_a$ is a binary relation on set I for agent $a \in \mathcal{A}$ such that*
   (a) *$\sim_a \subseteq \equiv_a$,*
   (b) *for each initial states $\alpha, \alpha', \alpha'' \in I$, if $\alpha \equiv_a \alpha'$ and $\alpha' \sim_a \alpha''$, then $\alpha \sim_a \alpha''$.*
6. *P is an arbitrary set of tuples $(\alpha, \delta, \omega) \in I \times \Delta^{\mathcal{A}} \times \Omega$ which we call the set of "plays",*
7. *$\pi(p) \subseteq P$ for each propositional variable p.*

As artificial agents are designed to be rational, to model the interplay between an agent's knowledge and cognitive ability, we require that an agent's cognition and knowledge be consistent. For example, if an agent knows that the sign is a stop sign, the agent should not recognize it as some other sign such as a no-entry sign. Such consistency is specified in Definition 1 item 5(a) and item 5(b).

Note that the relationship between the cognition relation and the equivalence relation for knowledge defined in item (5) resembles the relationship between the belief relation and the equivalence relation for knowledge defined by [24]. This is because one's cognition is tightly related to one's belief. However, cognition and belief are not the same concepts. Belief is about the states of a mind while cognition is the process of acquiring information. For autonomous agents, such a process is often through sensors, which can sometimes fail. For example, a Google Nest Doorbell will fail to recognize a delivery person in cold weather (when the temperature drops below $-4°$F) [14]. Thus, different from the belief relation in [24], our cognition relation need not be serial.

Intuitively, cognition in our work may sound related to the awareness in the Awareness Logic [10, 11, 26]. However, they are different concepts. Awareness Logic distinguishes two kinds of beliefs: implicit belief and explicit belief. The implicit belief is the same as the belief in [24]. It is called "implicit" because a consequence of what an agent believes might not be explicitly appreciated by the agent. Thus, an agent might believe in something deduced by the logic without being aware of it. From the standpoint that agents should not have explicit beliefs about propositions they are unaware of, Awareness Logic models awareness as a set of propositions for each state, where an agent's explicit beliefs are what the agent implicitly believes and is also aware of [10, 26]. Since our cognition relation differs from the (implicit) belief relation, as argued in the previous paragraph, our cognition is also distinct from the awareness in the Awareness Logic.

We consider non-deterministic situations when a complete action profile may lead to different outcomes. Thus, a play is a triple (α, δ, ω) rather than a pair (α, δ), see Definition 1 item (6).

Definition 2. *For any play (α, δ, ω) of a game $(I, \equiv, \Delta, \Omega, \sim, P, \pi)$ and any formula $\varphi \in \Phi$, the satisfaction relation $(\alpha, \delta, \omega) \Vdash \varphi$ is defined recursively:*

1. $(\alpha, \delta, \omega) \Vdash p$ *if* $(\alpha, \delta, \omega) \in \pi(p)$,
2. $(\alpha, \delta, \omega) \Vdash \neg\varphi$ *if* $(\alpha, \delta, \omega) \nVdash \varphi$,
3. $(\alpha, \delta, \omega) \Vdash \varphi \to \psi$ *if* $(\alpha, \delta, \omega) \nVdash \varphi$ *or* $(\alpha, \delta, \omega) \Vdash \psi$,
4. $(\alpha, \delta, \omega) \Vdash \mathsf{K}_a\varphi$ *if* $(\alpha', \delta', \omega') \Vdash \varphi$ *for each play* $(\alpha', \delta', \omega') \in P$ *such that* $\alpha \equiv_a \alpha'$,
5. $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a\varphi$ *when there is an action* $d \in \Delta$ *such that for each play* $(\alpha', \delta', \omega') \in P$, *if* $\alpha \sim_a \alpha'$ *and* $\delta'(a) = d$, *then* $(\alpha', \delta', \omega') \Vdash \varphi$.

The satisfaction relation $\Vdash$ in our semantics is a binary relation between a play and a formula. This approach is motivated by the intended meaning of the modality A, which represents an agent's ability to perform an action leading to a specific outcome. A play specifies a valid transition from an initial state to an outcome through an action profile. The modality A expresses that, under agent a's cognition, the agent has an action d that results in the outcome φ (see item (5) in the definition above). Therefore, a formula is a statement about a play, with propositional variables also representing statements about plays. Of course, such a statement could refer to the initial state, an action profile, or an outcome.

In item (5), the action $\delta'(a)$ is a uniform action under agent a's cognition. It is effective in all initial states α' such that $\alpha \sim_a \alpha'$. Such a uniform strategy has been studied in several logical systems where the focus was often the concept involving knowledge and action [2, 16, 32, 33, 45]. Different from these systems, our logical system mainly considers a uniform strategy for an agent's cognition to model the agent's cognitive ability.

As discussed before, the knowledge operator K represents the *ex ante* knowledge of an agent, as shown in item (4) of Definition 2, where the indistinguishability relation is applied to initial states. Moreover, for an intelligent agent to function effectively, it often relies on foundational information, such as traffic rules, stored in its knowledge base. The modality K also captures such knowledge, which is independent of the agent's actions.

4 AXIOMS AND THE SOUNDNESS

Our logical system contains all the propositional tautologies in language Φ and the following axioms.

1. Truth: $\mathsf{K}_a\varphi \to \varphi$,
2. Negative Introspection: $\neg\mathsf{K}_a\varphi \to \mathsf{K}_a\neg\mathsf{K}_a\varphi$,
3. K-Distributivity: $\mathsf{K}_a(\varphi \to \psi) \to (\mathsf{K}_a\varphi \to \mathsf{K}_a\psi)$,
4. A-Distributivity: $\mathsf{K}_a(\varphi \to \psi) \to (\mathsf{A}_a\varphi \to \mathsf{A}_a\psi)$,
5. Knowledge of Ability: $\mathsf{A}_a\varphi \to \mathsf{K}_a\mathsf{A}_a\varphi$.
6. Ability: $\neg\mathsf{A}_a\bot$.

The Truth, the Negative Introspection, and the K- Distributivity axioms are standard axioms from the epistemic logic S5. The A-Distributivity axiom states that if an agent knows that $\varphi \to \psi$ and the agent has a cognitive ability to provide information for the action leading to φ, then the agent has the ability to provide information for the action leading to ψ. Note that statement (2), which we discussed in Section 1, is an instance of the A-Distributivity axiom. The Knowledge of Ability axiom states that if an agent has a cognitive ability to provide information for its action that leads to an outcome, then the agent knows that it has such an ability. Statement (1), which we discussed in Section 1, is an instance of this axiom.

For the Ability axiom, recall that an autonomous agent may fail to acquire information due to extreme circumstances, causing the system to fail or terminate. In such cases, agents lose their cognitive ability. Since we study cognitive ability, we do not consider situations where the system terminates and an agent no longer has the cognitive ability. This is captured in the Ability axiom. Note that, by Definition 2 item (5), statement $\mathsf{A}_a\bot$ means that there is an action under the agent's cognition that leads to the outcome $\bot$. As $\bot$ is unsatisfiable, statement $\mathsf{A}_a\bot$ essentially indicates agent a's cognitive *inability*. Therefore, $\neg\mathsf{A}_a\bot$ represents that the agent possesses the cognitive ability.

Same as a belief modality B in [24], the modality A does not have a truth axiom. Thus, $\mathsf{A}_a\varphi$ may not imply φ. Intuitively, this is because an agent's cognition is limited. For example, in the Tesla accident where the car did not stop in front of the school bus stop sign and struck a teenager, the Autopilot considered that it was safe to drive and continued driving, denoted by A_{auto}"it is safe to drive". Unfortunately, it turned out that it was not safe to drive.

Note that modality A is not distributive, whereas a belief modality B in [24] is, written as $\mathsf{B}_a(\varphi \to \psi) \to (\mathsf{B}_a\varphi \to \mathsf{B}_a\psi)$. This is because modality A is meant to capture the cognitive ability to provide information about an action for an outcome rather than just capturing an agent's cognition. For example, consider the situation when an agent has a cognitive ability to foresee that an action, say d_1, will lead to $\varphi \to \psi$, and the agent also has the cognitive ability to foresee that an action, say d_2, will lead to φ. This does not mean that the agent has the cognitive ability to foresee an action that will lead to ψ. This is because action d_1 may be different from action d_2 and the agent may not have an action that would lead to ψ.

We say that a formula $\varphi \in \Phi$ is a *theorem* of our logical system, written as $\vdash \varphi$, if φ is derivable from the above axioms using the Modus Ponens and the Necessitation inference rules:

$$\frac{\varphi, \varphi \to \psi}{\psi}, \qquad \frac{\varphi}{\mathsf{K}_a\varphi}.$$

We write $X \vdash \varphi$ if a formula $\varphi \in \Phi$ is derivable from the *theorems* of our logical system and an additional set of assumptions $X \subseteq \Phi$ using only the Modus Ponens inference rule. A set X is said to be consistent if $X \nvdash \bot$.

The soundness of propositional tautologies and of the Modus Ponens and the Necessitation inference rules is straightforward. The soundness of the Truth axiom, the Negative Introspection axiom, and the K-Distributivity axiom is standard. Below we prove the soundness of the axioms related to the modality A. Let $\varphi \in \Phi$ and $(\alpha, \delta, \omega) \in P$ be a play of a game $(I, \equiv, \Delta, \Omega, \sim, P, \pi)$.

Lemma 4.1. *If $(\alpha, \delta, \omega) \Vdash \mathsf{K}_a(\varphi \to \psi)$ and $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a\varphi$, then $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a\psi$.*

Proof. The assumption $(\alpha, \delta, \omega) \Vdash \mathsf{K}_a(\varphi \to \psi)$, by Definition 2 item (4), implies that for each play $(\alpha_1, \delta_1, \omega_1)$ such that $\alpha \equiv_a \alpha'$, we have

$$(\alpha_1, \delta_1, \omega_1) \Vdash \varphi \to \psi. \tag{3}$$

The assumption $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a\varphi$, by Definition 2 item (5), implies that there is an action $d \in \Delta$ such that for each play $(\alpha_2, \delta_2, \omega_2) \in P$,

$$\text{if } \alpha \sim_a \alpha_2 \text{ and } \delta_2(a) = d, \text{ then } (\alpha_2, \delta_2, \omega_2) \Vdash \varphi. \tag{4}$$

Consider an arbitrary play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_a \alpha'$ and $\delta'(a) = d$. Then, $(\alpha', \delta', \omega') \Vdash \varphi$ by statement (4). At the same time, $\alpha \equiv_a \alpha'$ by Definition 1 item 5(a). Thus, it follows from statement (3) that $(\alpha', \delta', \omega') \Vdash \psi$. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a\psi$, by Definition 2 item (5). □

Lemma 4.2. *If $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a\varphi$, then $(\alpha, \delta, \omega) \Vdash \mathsf{K}_a\mathsf{A}_a\varphi$.*

Proof. Assumption $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a\varphi$ implies that there is an action $d_0 \in \Delta$ such that for each play $(\alpha', \delta', \omega') \in P$,

$$\text{if } \alpha \sim_a \alpha' \text{ and } \delta'(a) = d_0, \text{ then } (\alpha', \delta', \omega') \Vdash \varphi. \tag{5}$$

Consider any play $(\alpha', \delta', \omega') \in P$ such that $\alpha \equiv_a \alpha'$. By Definition 2 item (4), it suffices to show that $(\alpha', \delta', \omega') \Vdash \mathsf{A}_a\varphi$. That is, we need to show that there is an action $d \in \Delta$ such that for each play $(\alpha'', \delta'', \omega'') \in P$, if $\alpha' \sim_a \alpha''$ and $\delta''(a) = d$, then $(\alpha'', \delta'', \omega'') \Vdash \varphi$. Let d be d_0. Note that assumptions $\alpha \equiv_a \alpha'$ and $\alpha' \sim_a \alpha''$ imply that $\alpha \sim_a \alpha''$ by Definition 1 item 5(b). Then, $(\alpha'', \delta'', \omega'') \Vdash \varphi$ by statement (5). □

Lemma 4.3. $(\alpha, \delta, \omega) \nVdash \mathsf{A}_a\bot$.

Proof. Suppose that $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a\bot$. Then, By Definition 2 item (5), there is an action $d \in \Delta$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_a \alpha'$ and $\delta'(a) = d$, then $(\alpha', \delta', \omega') \Vdash \bot$. However, consider any play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_a \alpha'$ and $\delta'(a) = d$. It follows that $(\alpha', \delta', \omega') \nVdash \bot$, which leads to a contradiction. □

Next, we list five lemmas that will be used later in the proof of completeness. The proofs of the first three lemmas can be found in the Appendix section. Lemma 4.4 is the well-known positive introspection principle. Lemma 4.7 says that if an agent does not have the cognitive ability to perform an action that leads to an outcome, then the agent knows that he does not have such an ability. For the Lindenbaum's lemma (Lemma 4.8), the standard proof applies (see, e.g. Mendelson [31, Proposition 2.14]).

Lemma 4.4. $\vdash \mathsf{K}_a\varphi \to \mathsf{K}_a\mathsf{K}_a\varphi$.

Lemma 4.5 (deduction). *If $X, \varphi \vdash \psi$, then $X \vdash \varphi \to \psi$.*

Lemma 4.6. *If $\varphi_1 \wedge \cdots \wedge \varphi_n \vdash \psi$, then $\mathsf{K}_a\varphi_1, \dots, \mathsf{K}_a\varphi_n \vdash \mathsf{K}_a\psi$.*

Lemma 4.7. $\vdash \neg\mathsf{A}_a\varphi \to \mathsf{K}_a\neg\mathsf{A}_a\varphi$.

Proof. By the Knowledge of Ability axiom, $\vdash \mathsf{A}_a\varphi \to \mathsf{K}_a\mathsf{A}_a\varphi$. Thus, $\vdash \neg\mathsf{K}_a\mathsf{A}_a\varphi \to \neg\mathsf{A}_a\varphi$ by the contrapositive. Hence, by the Necessitation inference rule, $\vdash \mathsf{K}_a(\neg\mathsf{K}_a\mathsf{A}_a\varphi \to \neg\mathsf{A}_a\varphi)$. Then, by the Distributivity axiom and the Modus Ponens inference rule $\vdash \mathsf{K}_a\neg\mathsf{K}_a\mathsf{A}_a\varphi \to \mathsf{K}_a\neg\mathsf{A}_a\varphi$. Thus, $\vdash \neg\mathsf{K}_a\mathsf{A}_a\varphi \to \mathsf{K}_a\neg\mathsf{A}_a\varphi$, by the Negative Introspection axiom and the laws of propositional reasoning. Note that $\neg\mathsf{A}_a\varphi \to \neg\mathsf{K}_a\mathsf{A}_a\varphi$ is the contrapositive of the Truth axiom. Therefore, by the laws of propositional reasoning, $\vdash \neg\mathsf{A}_a\varphi \to \mathsf{K}_a\neg\mathsf{A}_a\varphi$. □

Lemma 4.8 (Lindenbaum). *Any consistent set of formulae can be extended to a maximal consistent set of formulae.*

5 COMPLETENESS

In this section, we prove the completeness of our logical system, listed as Theorem 5.1 below.

Theorem 5.1. *If $X \nvdash \varphi$, then there is a game and a play (α, δ, ω) in the game such that $(\alpha, \delta, \omega) \Vdash \chi$ for each $\chi \in X$ and $(\alpha, \delta, \omega) \nVdash \varphi$.*

Towards the proof of this theorem, we will construct a game, called the canonical game, such that a play in the game satisfies all formulas in X, but does not satisfy formula φ. We use the tuple $(I, \equiv, \Delta, \Omega, \sim, P, \pi)$ to denote the canonical game G. The following definitions, Definition 3 to Definition 11, specify each component of the canonical game G.

Definition 3. *The set of outcomes Ω is the set of all maximal consistent sets of formulae.*

Definition 4. *For each $\omega_1, \omega_2 \in \Omega$, we say that $\omega_1 \equiv_a \omega_2$ when, for each $\varphi \in \Phi$, if $\mathsf{K}_a\varphi \in \omega_1$, then $\mathsf{K}_a\varphi \in \omega_2$.*

Lemma 5.2. *If $\omega_1 \equiv_a \omega_2$, then for each $\varphi \in \Phi$, $\mathsf{K}_a\varphi \in \omega_1$ if and only if $\mathsf{K}_a\varphi \in \omega_2$.*

Proof. ($\Rightarrow$) By Definition 4, if $\mathsf{K}_a\varphi \in \omega_1$, then $\mathsf{K}_a\varphi \in \omega_2$.

($\Leftarrow$) Suppose that $\mathsf{K}_a\varphi \notin \omega_1$. Then, $\neg\mathsf{K}_a\varphi \in \omega_1$ because ω_1 is maximal. Thus, $\omega_1 \vdash \mathsf{K}_a\neg\mathsf{K}_a\varphi$ by the Negative Introspection axiom. Hence, $\mathsf{K}_a\neg\mathsf{K}_a\varphi \in \omega_1$ since ω_1 is maximal. Then, by Definition 4, it follows from the assumption $\omega_1 \equiv_a \omega_2$ that $\mathsf{K}_a\neg\mathsf{K}_a\varphi \in \omega_2$. Thus, $\omega_2 \vdash \neg\mathsf{K}_a\varphi$ by the Truth axiom. Therefore, $\neg\mathsf{K}_a\varphi \in \omega_2$ because ω_2 is maximal. □

The next lemma directly follows from Lemma 5.2.

Lemma 5.3. *Relation $\equiv_a$ is an equivalence relation.*

Definition 5. $\omega_1 \equiv_{\mathcal{A}} \omega_2$ *if* $\omega_1 \equiv_a \omega_2$ *for every* $a \in \mathcal{A}$.

The set I of initial states is the set of equivalence classes with respect to the relation $\equiv_{\mathcal{A}}$.

Definition 6. $I := \Omega/\equiv_{\mathcal{A}}$.

Lemma 5.4. *Relation $\equiv_a$ is well-defined on set I.*

Proof. Suppose that $\omega_1 \equiv_a \omega_2$. Consider any outcomes ω_1' and ω_2' such that $\omega_1 \equiv_{\mathcal{A}} \omega_1'$ and $\omega_2 \equiv_{\mathcal{A}} \omega_2'$. It suffices to prove that $\omega_1' \equiv_a \omega_2'$. Note that $\mathsf{K}_a\varphi \in \omega_1$ if and only if $\mathsf{K}_a\varphi \in \omega_1'$ by the assumption $\omega_1 \equiv_{\mathcal{A}} \omega_1'$, Definition 5, and Lemma 5.2. Similarly, it follows from the assumption $\omega_2 \equiv_{\mathcal{A}} \omega_2'$ that $\mathsf{K}_a\varphi \in \omega_2$ if and only if $\mathsf{K}_a\varphi \in \omega_2'$. Since $\omega_1 \equiv_a \omega_2$, we have $\mathsf{K}_a\varphi \in \omega_1$ if and only if $\mathsf{K}_a\varphi \in \omega_2$, again by Lemma 5.2. Thus, $\mathsf{K}_a\varphi \in \omega_1'$ if and only if $\mathsf{K}_a\varphi \in \omega_2'$ by propositional reasoning. Therefore, $\omega_1' \equiv_a \omega_2'$ by Definition 4. □

The next lemma directly follows from Lemma 5.4.

Lemma 5.5. $\alpha \equiv_a \alpha'$ *if and only if* $\omega \equiv_a \omega'$, *for any initial states* $\alpha, \alpha' \in I$ *and any outcomes* $\omega \in \alpha$ *and* $\omega' \in \alpha'$.

Note that an initial state is an equivalence class of outcomes. The next lemma shows that an agent maintains the same cognitive ability in the outcomes that belong to the same initial state.

Lemma 5.6. $\mathsf{A}_a\varphi \in \omega$ *if and only if* $\mathsf{A}_a\varphi \in \omega'$, *for each* $\omega, \omega' \in \Omega$ *such that* $\omega \equiv_a \omega'$ *and each* $\varphi \in \Phi$.

Proof. Since relation $\equiv_a$ is an equivalence relation by Lemma 5.3, it suffices to show that if $\mathsf{A}_a\varphi \in \omega$, then $\mathsf{A}_a\varphi \in \omega'$. Assume $\mathsf{A}_a\varphi \in \omega$. Then, by the Knowledge of Ability axiom, $\omega \vdash \mathsf{K}_a\mathsf{A}_a\varphi$. Hence, $\mathsf{K}_a\mathsf{A}_a\varphi \in \omega$ because ω is a maximal consistent set. Thus, $\mathsf{K}_a\mathsf{A}_a\varphi \in \omega'$ by the statement $\omega \equiv_a \omega'$ and Lemma 5.2. Hence, $\omega' \vdash \mathsf{A}_a\varphi$ by the Truth axiom and the Modus Ponens rule. Therefore, $\mathsf{A}_a\varphi \in \omega'$ since ω' is a maximal consistent set. □

To define the cognitive relation $\sim$ that satisfies the conditions in item (5) of Definition 1, we must ensure that the relation $\sim_a$ is a subset of $\equiv_a$. Additionally, recall that an intelligent agent may lose its cognitive ability in extreme circumstances, and that $\neg\mathsf{A}_a\bot$ indicates that agent a retains cognitive ability. Therefore, the cognitive relation only considers pairs of states where the agent possesses cognitive ability, as outlined in the definition below.

Definition 7. $\sim_a := \{(\omega_1, \omega_2) \mid \omega_1 \equiv_a \omega_2$ *and* $\neg\mathsf{A}_a\bot \in \omega_1\}$.

Definition 7 defines the cognition relation on the set of outcomes and the relation $\sim_a$ is a proper subset of the relation $\equiv_a$ for any agent $a \in \mathcal{A}$. This definition is used to define the cognition relation on the set of initial states, as shown in the next definition.

Definition 8. $\alpha_1 \sim_a \alpha_2$ *if* $\alpha_1 \equiv_a \alpha_2$ *and for each* $\omega_1 \in \alpha_1$, *there is* $\omega_2 \in \alpha_2$ *such that* $\omega_1 \sim_a \omega_2$.

To show that the relation $\sim_a$ on set I is well-defined, we need to prove that it satisfies conditions 5(a) and 5(b) in Definition 1.

Lemma 5.7. *Relation $\sim_a$ on set I satisfies item* (5) *of Definition 1.*

Proof. We need to show that relation $\sim_a$ satisfies both condition (a) and condition (b) in item (5) of Definition 1.

Condition (a): $\sim_a \subseteq \equiv_a$ holds true by Definition 7.

For condition (b), consider arbitrary initial states $\alpha_1, \alpha_2, \alpha_3 \in I$. Assume $\alpha_1 \equiv_a \alpha_2$ and $\alpha_2 \sim_a \alpha_3$. It suffices to show $\alpha_1 \sim_a \alpha_3$.

Assumption $\alpha_2 \sim_a \alpha_3$, by Definition 8, implies that $\alpha_2 \equiv_a \alpha_3$. Then, by Lemma 5.5, Lemma 5.3, and the assumption $\alpha_1 \equiv_a \alpha_2$, we have $\alpha_1 \equiv_a \alpha_3$ (∗). Thus, to show $\alpha_1 \sim_a \alpha_3$, by Definition 8, we need to show that for each $\omega_1 \in \alpha_1$, there is $\omega_3 \in \alpha_3$ such that $\omega_1 \sim_a \omega_3$.

Consider any $\omega_1 \in \alpha_1$ and any $\omega_3 \in \alpha_3$. Then, statement (∗) implies $\omega_1 \equiv_a \omega_3$, by Lemma 5.5. Thus, by Definition 7, it suffices to show that $\neg\mathsf{A}_a\bot \in \omega_1$.

Consider any $\omega_2 \in \alpha_2$. Assumption $\alpha_2 \sim_a \alpha_3$, by Definition 8 and Definition 7, implies that $\neg\mathsf{A}_a\bot \in \omega_2$. Moreover, assumption $\alpha_1 \equiv_a \alpha_2$ implies $\omega_1 \equiv_a \omega_2$ by Lemma 5.5. Thus, $\neg\mathsf{A}_a\bot \in \omega_1$ by Lemma 5.6. Therefore, the desired is true. □

The set of actions is defined to be the set of all formulae.

Definition 9. $\Delta = \Phi$.

The set P of valid plays is defined in the next definition. Intuitively, from any initial state, agents take actions that lead to an outcome. By Definition 2 item (5), an agent's ability to provide information about an action for the outcome φ is interpreted as the agent having an action such that, under the agent's cognition, this action will lead to φ. As a result, φ should hold true in the outcome. In the canonical model, we use φ itself as the agent's action, see Definition 10 below.

Next, recall that $\neg\mathsf{A}_a\bot$ represents that the agent possesses the cognitive ability. When an autonomous agent loses its cognitive ability, the system may terminate. Since a valid play represents a system transition, it should not include situations when system fails and an agent no longer has its cognitive ability. Thus, $\neg\mathsf{A}_a\bot$ is in the outcome for valid plays.

In Definition 6, initial states are defined to be the equivalence classes with respect to the relation $\equiv_{\mathcal{A}}$. This is to ensure that conditions required for valid plays work for any agent in set $\mathcal{A}$. The next definition specifies the set of valid plays.

Definition 10. *The set* $P \subseteq I \times \Delta^{\mathcal{A}} \times \Omega$ *consists of all triples* (α, δ, ω) *such that* $\omega \in \alpha$, $\neg\mathsf{A}_a\bot \in \omega$ *for each* $a \in \mathcal{A}$, *and for each* $\mathsf{A}_a\varphi \in \omega$, *if* $\delta(a) = \varphi$, *then* $\varphi \in \omega$.

Definition 11. $\pi(p) = \{(\alpha, \delta, \omega) \in P \mid p \in \omega\}$ *for any atomic proposition* p.

This concludes the definition of the canonical game G. The next lemma shows that given any outcome that contains $\neg\mathsf{A}_a\bot$, a play can always be constructed from the outcome.

Lemma 5.8. *For any outcome* $\omega \in \Omega$ *where* $\neg\mathsf{A}_a\bot \in \omega$, *there is an initial state* $\alpha \in I$ *and a complete action profile* $\delta \in \Delta^{\mathcal{A}}$ *such that* $(\alpha, \delta, \omega) \in P$.

Proof. Let α be the equivalence class of ω with respect to relation $\equiv_{\mathcal{A}}$. Thus, $\omega \in \alpha$. Let $\delta(a) = \top$ for each agent $a \in \mathcal{A}$. Consider any formula $\mathsf{A}_a\varphi \in \omega$ such that $\delta(a) = \varphi$. By Definition 10 and the assumption $\neg\mathsf{A}_a\bot \in \omega$, it suffices to show that $\varphi \in \omega$. Indeed, since

$\delta(a) = \top$, by the assumption $\delta(a) = \varphi$, we have $\varphi = \top$. Therefore, $\varphi \in \omega$ because set ω is maximal. □

In proving the completeness theorem, a key step of the proof is an induction lemma, so-called the "truth lemma" (Lemma 5.13 in this paper). This lemma is proven by induction on the structure complexity of the language. The following four lemmas address the sub-cases within the induction step for Lemma 5.13, specifically for the two modalities: the knowledge modality K and the cognitive ability modality A.

Lemma 5.9. *For any play $(\alpha, \delta, \omega) \in P$, if $\mathsf{K}_a\varphi \in \omega$, then $\varphi \in \omega'$ for each play $(\alpha', \delta', \omega') \in P$ such that $\alpha \equiv_a \alpha'$.*

Proof. By Definition 10, assumption $(\alpha, \delta, \omega) \in P$ implies $\omega \in \alpha$. Consider an arbitrary play $(\alpha', \delta', \omega') \in P$ such that $\alpha \equiv_a \alpha'$. Then, $\omega' \in \alpha'$ by Definition 10. By Lemma 5.5 and assumption $\alpha \equiv_a \alpha'$, we have $\omega \equiv_a \omega'$. Thus, $\mathsf{K}_a\varphi \in \omega'$ by Definition 4 and the assumption $\mathsf{K}_a\varphi \in \omega$ of the lemma. Thus, $\omega' \vdash \varphi$ by the Truth axiom and the Modus Ponens rule. Therefore, $\varphi \in \omega'$ because ω' is maximal by Definition 3. □

Lemma 5.10. *For any play $(\alpha, \delta, \omega) \in P$, if $\mathsf{K}_a\varphi \notin \omega$, then there exists a play $(\alpha', \delta', \omega') \in P$ such that $\alpha \equiv_a \alpha'$ and $\varphi \notin \omega'$.*

Proof. Consider set $X := \{\neg\varphi\} \cup \{\chi \mid \mathsf{K}_a\chi \in \omega\}$.

Claim 1. *Set X is consistent.*

Proof of Claim. Suppose the opposite. Then, there are formulas

$$\mathsf{K}_a\chi_1, \mathsf{K}_a\chi_2, \dots, \mathsf{K}_a\chi_n \in \omega \tag{6}$$

such that $\chi_1 \wedge \chi_2 \wedge \dots \wedge \chi_n \vdash \varphi$. Then, $\mathsf{K}_a\chi_1, \mathsf{K}_a\chi_2, \dots, \mathsf{K}_a\chi_n \vdash \mathsf{K}_a\varphi$, by Lemma 4.6. Thus, by statement (6), we have $\omega \vdash \mathsf{K}_a\varphi$, which contradicts the assumption that $\mathsf{K}_a\varphi \notin \omega$ because set ω is a maximal consistent set. Therefore, set X is consistent. ⊠

Let ω' be a maximal consistent extension of set X. By Lemma 4.8, ω' exists. Let α' be the equivalence class of ω'. Thus, $\omega' \in \alpha'$.

Claim 2. $\omega \equiv_a \omega'$.

Proof of Claim. By Definition 4, we need to prove that for each $\tau \in \Phi$, if $\mathsf{K}_a\tau \in \omega$, then $\mathsf{K}_a\tau \in \omega'$. Suppose $\mathsf{K}_a\tau \in \omega$. Then, $\omega \vdash \mathsf{K}_a\mathsf{K}_a\tau$ by Lemma 4.4 and the Modus Ponens rule. Hence, $\mathsf{K}_a\mathsf{K}_a\tau \in \omega$ because ω is maximal. Thus, $\mathsf{K}_a\tau \in X \subseteq \omega'$, by the definition of set X and the choice of ω'. ⊠

Next, we define the complete action profile δ'. For each agent $b \in \mathcal{A}$, let

$$\delta'(b) = \top. \tag{7}$$

Claim 3. $(\alpha', \delta', \omega') \in P$.

Proof of Claim. First, note that $\omega' \in \alpha'$. Next, we show that $\neg\mathsf{A}_a\bot \in \omega'$. Indeed, assumption $(\alpha, \delta, \omega) \in P$ implies $\neg\mathsf{A}_a\bot \in \omega$ by Definition 10. Thus, $\neg\mathsf{A}_a\bot \in \omega'$ by Claim 2 and Lemma 5.6. Finally, consider any $\mathsf{A}_b\psi \in \omega'$ such that $\delta'(b) = \psi$. By Definition 10, it suffices to show that $\psi \in \omega'$. Indeed, $\delta'(b) = \top$ by equation (7). Hence, $\psi = \delta'(b) = \top$. Therefore, $\psi \in \omega'$ because set ω' is maximal. ⊠

To complete the proof, note that $\alpha \equiv_a \alpha'$ by Lemma 5.5 and Claim 2. Also notice $\neg\varphi \in X \subseteq \omega'$ by the definition of set X and the choice of set ω'. Therefore, $\varphi \notin \omega'$ because ω' is a maximal consistent set. □

Lemma 5.11. *For any play $(\alpha, \delta, \omega) \in P$ and $\mathsf{A}_a\varphi \in \omega$, there is an action $d \in \Delta$ such that for each play $(\alpha', \delta', \omega') \in P$ where $\alpha \sim_a \alpha'$ and $\delta'(a) = d$, we have $\varphi \in \omega'$.*

Proof. Let $d = \varphi$. Consider any play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_a \alpha'$ and $\delta'(a) = d$. Then, $\alpha \equiv_a \alpha'$ by Definition 8. By Definition 10, assumption $(\alpha', \delta', \omega') \in P$ implies that $\omega' \in \alpha'$. Similarly, assumption $(\alpha, \delta, \omega) \in P$ implies that $\omega \in \alpha$. Then, it follows from Lemma 5.5 that $\omega \equiv_a \omega'$. Thus, assumption $\mathsf{A}_a\varphi \in \omega$, by Lemma 5.6, implies $\mathsf{A}_a\varphi \in \omega'$. Therefore, $\varphi \in \omega'$ by the assumption $(\alpha', \delta', \omega') \in P$, the assumption $\delta'(a) = d = \varphi$, and Definition 10. □

Lemma 5.12. *For any play $(\alpha, \delta, \omega) \in P$, if $\mathsf{A}_a\varphi \notin \omega$, then for each $d \in \Delta$, there exists a play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_a \alpha'$, $\delta'(a) = d$, and $\varphi \notin \omega'$.*

Proof. Consider set

$$X := \{\neg\varphi\} \cup \{\psi \mid \mathsf{A}_a\psi \in \omega, d = \psi\} \cup \{\chi \mid \mathsf{K}_a\chi \in \omega\}.$$

Claim 4. *Set X is consistent.*

Proof of Claim. Suppose the opposite. Then there are formulas

$$\mathsf{A}_a\psi, \mathsf{K}_a\chi_1, \mathsf{K}_a\chi_2, \dots, \mathsf{K}_a\chi_n \in \omega \tag{8}$$

$$\text{such that} \qquad d = \psi \tag{9}$$

$$\text{and} \qquad \psi \wedge \chi_1 \wedge \dots \wedge \chi_n \vdash \varphi.$$

By Lemma 4.5, $\chi_1 \wedge \dots \wedge \chi_n \vdash \psi \to \varphi$. Then, by Lemma 4.6, $\mathsf{K}_a\chi_1, \dots, \mathsf{K}_a\chi_n \vdash \mathsf{K}_a(\psi \to \varphi)$. By the A-Distributivity axiom and the Modus Ponens rule, $\mathsf{K}_a\chi_1, \dots, \mathsf{K}_a\chi_n \vdash \mathsf{A}_a\psi \to \mathsf{A}_a\varphi$. Hence, $\mathsf{K}_a\chi_1, \mathsf{K}_a\chi_2, \dots, \mathsf{K}_a\chi_n, \mathsf{A}_a\psi \vdash \mathsf{A}_a\varphi$, by the Modus Ponens inference rule. Thus, $\omega \vdash \mathsf{A}_a\varphi$, by statement (8). This contradicts the assumption that $\mathsf{A}_a\varphi \notin \omega$ because set ω is a maximal consistent set. Therefore, set X is consistent. ⊠

Let ω' be a maximal consistent extension of set X. By Lemma 4.8, ω' exists. Let α' be the equivalence class of ω'. Thus, $\omega' \in \alpha'$.

Claim 5. $\omega \equiv_a \omega'$.

Proof of Claim. By Definition 4, it suffices to show that if $\mathsf{K}_a\tau \in \omega$ then $\mathsf{K}_a\tau \in \omega'$, for each $\tau \in \Phi$. Assume that $\mathsf{K}_a\tau \in \omega$. Then, $\omega \vdash \mathsf{K}_a\mathsf{K}_a\tau$, by Lemma 4.4. Since ω is maximal, $\mathsf{K}_a\mathsf{K}_a\tau \in \omega$. Thus, by the definition of set X and the choice of ω', $\mathsf{K}_a\tau \in X \subseteq \omega'$. ⊠

Claim 6. $\alpha \sim_a \alpha'$.

Proof of Claim. First, note that $\alpha \equiv_a \alpha'$ by Lemma 5.5 and Claim 5. Then, by Definition 8, it suffices to show that for each $\omega_1 \in \alpha$, there is $\omega_2 \in \alpha'$ such that $\omega_1 \sim_a \omega_2$. Choose an arbitrary $\omega_1 \in \alpha$. For any $\omega_2 \in \alpha'$, we have $\omega_1 \equiv_a \omega_2$ by Lemma 5.5 and the statement $\alpha \equiv_a \alpha'$. Then, by Definition 7, it suffices to show that $\neg\mathsf{A}_a\bot \in \omega_1$. Note that assumption $(\alpha, \delta, \omega) \in P$ implies that $\omega \in \alpha$ and $\neg\mathsf{A}_a\bot \in \omega$ by Definition 10. Thus, $\omega \equiv_a \omega_1$ by Definition 6 and the assumption $\omega_1 \in \alpha$. Hence, $\neg\mathsf{A}_a\bot \in \omega_1$ by Lemma 5.6. ⊠

Next, we define the complete action profile δ'. For each agent $b \in \mathcal{A}$, let

$$\delta'(b) := \begin{cases} d, & \text{if } b = a, \\ \top, & \text{otherwise.} \end{cases} \tag{10}$$

Claim 7. $(\alpha', \delta', \omega') \in P$.

Proof of Claim. First, note that $\omega' \in \alpha'$ by the definition of α'. Also note that $\neg \mathsf{A}_a \bot \in \omega$ by the assumption $(\alpha, \delta, \omega) \in P$ and Definition 10. Thus, $\neg \mathsf{A}_a \bot \in \omega'$ by Claim 5 and Lemma 5.6. Finally, consider any $\mathsf{A}_b \psi \in \omega'$ such that $\delta'(b) = \psi$. It suffices to show that $\psi \in \omega'$ by Definition 10.

Case I: $b = a$. Then $\mathsf{A}_a \psi \in \omega$ by the assumption $\mathsf{A}_b \psi \in \omega'$, the assumption $b = a$ of the case, Claim 5, and Lemma 5.6. At the same time, $d = \psi$ by equation (9). Therefore, we have $\psi \in X \subseteq \omega'$ by the definition of X and choice of ω'.

Case II: $b \neq a$. Then, $\delta'(b) = \top$ by equation (10). Hence, $\psi = \delta'(b) = \top$. Therefore, $\psi \in \omega'$ because set ω' is maximal. ⊠

To finish the proof, notice that $\alpha \sim_a \alpha'$ by Claim 6. Moreover, $\neg\varphi \in X \subseteq \omega'$. Thus, $\varphi \notin \omega'$ since ω' is consistent. □

The next lemma is the truth lemma in our completeness proof.

Lemma 5.13. *$(\alpha, \delta, \omega) \Vdash \varphi$ if and only if $\varphi \in \omega$.*

Proof. We prove the lemma by induction on the complexity of formula φ. If φ is a propositional variable, then the lemma follows from item (1) of Definition 2 and Definition 11. If formula φ is an implication or a negation, then the lemma follows from items (2) or (3) of Definition 2 and the fact that ω is a maximal consistent set in the standard way.

Suppose that formula φ is of the form $\mathsf{K}_a \varphi$.

($\Leftarrow$) Assume $\mathsf{K}_a \varphi \in w$. Consider any $(\alpha', \delta', \omega') \in P$ such that $\alpha \equiv_a \alpha'$. Then, by Lemma 5.9, we have $\varphi \in \omega'$. Thus, by the induction hypothesis, $(\alpha', \delta', \omega') \Vdash \varphi$. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{K}_a \varphi$ by Definition 2 item (4).

($\Rightarrow$) Assume $\mathsf{K}_a \varphi \notin w$. Then by Lemma 5.10, there is a play $(\alpha', \delta', \omega') \in P$ such that $\alpha \equiv_a \alpha'$ and $\varphi \notin \omega'$. Thus, by the induction hypothesis, $(\alpha', \delta', \omega') \nVdash \varphi$. Therefore, $(\alpha, \delta, \omega) \nVdash \mathsf{K}_a \varphi$ by Definition 2 item (4).

Suppose that formula φ is of the form $\mathsf{A}_a \varphi$.

($\Leftarrow$) Assume $\mathsf{A}_a \varphi \in w$. Then, by Lemma 5.11, there is an action $d \in \Delta$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_a \alpha'$ and $\delta'(a) = d$, then $\varphi \in \omega'$. Thus, by the induction hypothesis, $(\alpha', \delta', \omega') \Vdash \varphi$. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{A}_a \varphi$ by Definition 2 item (5).

($\Rightarrow$) Let $\mathsf{A}_a \varphi \notin \omega$. Then, $\neg \mathsf{A}_a \varphi \in \omega$ by the maximality of set ω. Hence, by Lemma 5.12, for each $d \in \Delta$, there is a play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_a \alpha'$, $\delta'(a) = d$, and $\neg\varphi \in \omega'$. Thus $\varphi \notin \omega'$ by the consistency of set ω'. Hence, by the induction hypothesis, $(\alpha', \delta', \omega') \nVdash \varphi$. Thus, $(\alpha, \delta, \omega) \nVdash \mathsf{A}_a \varphi$ by Definition 2 item (5). □

To finish the proof of completeness, assume that $X \nvdash \varphi$. We need to show that there is a game and a play (α, δ, ω) in the game such that $(\alpha, \delta, \omega) \Vdash \chi$ for each $\chi \in X$ and $(\alpha, \delta, \omega) \nVdash \varphi$. Since $X \nvdash \varphi$, the set $X \cup \{\neg\varphi\}$ is consistent. By Lemma 4.8, there is a maximal consistent extension ω_0 of the set $X \cup \{\neg\varphi\}$.

Consider the canonical game $G = (I, \{\equiv_a\}_{a \in \mathcal{A}}, \Delta, \Omega, \{\sim_a\}_{a \in \mathcal{A}}, P, \pi)$. By Lemma 5.3, relation $\equiv_a$ in the game G satisfies the condition in item (2) of Definition 1. By Lemma 5.7, relation $\sim_a$ satisfies conditions in item (5) of Definition 1. Thus, the canonical game G is indeed a game defined in Definition 1.

Note that $\omega_0 \in \Omega$ by Definition 3. Then, $\omega_0 \vdash \neg \mathsf{A}_a \bot$, by the Ability axiom. Thus, $\neg \mathsf{A}_a \bot \in \omega_0$, since ω_0 is a maximal consistent set. Hence, by Lemma 5.8, there is an initial state α_0 and a complete action profile δ_0 such that $(\alpha_0, \delta_0, \omega_0) \in P$. Since $X \cup \{\neg\varphi\} \subseteq \omega_0$, by Lemma 5.13, we have $(\alpha_0, \delta_0, \omega_0) \Vdash \chi$ for each $\chi \in X \subseteq \omega_0$ and $(\alpha_0, \delta_0, \omega_0) \Vdash \neg\varphi$. Thus, $(\alpha_0, \delta_0, \omega_0) \nVdash \varphi$.

6 CONCLUSION AND FUTURE WORK

In this paper, we explored the cognitive ability of intelligent agents and introduced a logical system to reason about the interplay between knowledge and cognitive ability of intelligent agents. In our framework, an agent's cognitive ability refers to its ability to perform an action leading to specific outcomes under its cognition. We argue that an agent's cognition differs from its belief and that must remain consistent with its knowledge. Finally, we provided a sound and complete axiomatization for our proposed system.

Our work is intended to be a foundational step toward in-depth studies of cognitive ability of intelligent agents that can be potentially used to recognize hazards, communicate with human operators before an unfavorable outcome, and assist retrospective analysis afterwards, thus enhancing safety and reliability of such systems. We do not consider complex structures of autonomous agents and their environments, such as the limited resources necessary for an agent to achieve its goal.

There are potentially many directions to extend the current work. For example, one direction is to extend the current work to account for dynamic environments where the knowledge base evolves through interactions with the environment. The history of actions taken need to be considered and the assumption of perfect recall for intelligent agents also is reasonable.

Another direction is to consider the collaboration of multiple agents, for example, the collaboration between the self-driving car and the backup driver. Traditionally, the concept of a coalition's knowledge extends individual knowledge to a group of agents. This includes distributed knowledge where members share their knowledge with each other, common knowledge, and group knowledge where each member individually knows the same thing. However, these notions of knowledge do not align well with the dynamics of collaboration between a self-driving car and a backup driver. Consider a situation when one agent recognizes a potential risk of collision. Then an action should be taken immediately to prevent the collision. This highlights the need for a new type of coalitional knowledge and effective collaboration may require a more nuanced understanding of knowledge and cognition ability.

REFERENCES

[1] Aldo Iván Ramírez Abarca and Jan M Broersen. 2022. A Stit Logic of Responsibility.. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1717–1719.

[2] Thomas Ågotnes, Valentin Goranko, Wojciech Jamroga, and Michael Wooldridge. 2015. Knowledge and ability. In *Handbook of epistemic logic*, Hans Van Ditmarsch, Wiebe van der Hoek, Joseph Y Halpern, and Barteld Kooi (Eds.). College Publications, Rickmansworth, UK, 543–589.

[3] Natasha Alechina, Brian Logan, Hoang Nga Nguyen, and Abdur Rakib. 2011. Logic for coalitions with bounded resources. *Journal of Logic and Computation* 21, 6 (2011), 907–937.

[4] Philippe Balbiani, Andreas Herzig, and Nicolas Troquard. 2008. Alternative axiomatics and complexity of deliberative STIT theories. *Journal of Philosophical Logic* 37 (2008), 387–406.

[5] National Transportation Safety Board. 2018. *Preliminary Report Highway HWY18MH010.* NTSB. https://web.archive.org/web/20190831200841/https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf

[6] Neal E. Boudette, Cade Metz, and Jack Ewing. 2022. *Tesla Autopilot and Other Driver-Assist Systems Linked to Hundreds of Crashes.* New York Times. https://www.nytimes.com/2022/06/15/business/self-driving-car-nhtsa-crash-data.html

[7] Jan Broersen. 2009. A Complete STIT Logic for Knowledge and Action, and Some of Its Applications. In *Declarative Agent Languages and Technologies VI*, Matteo Baldoni, Tran Cao Son, M. Birna van Riemsdijk, and Michael Winikoff (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 47–59.

[8] Jan M Broersen. 2011. Making a start with the stit logic analysis of intentional action. *Journal of philosophical logic* 40 (2011), 499–530.

[9] Mark A. Brown. 1988. On the Logic of Ability. *Journal of Philosophical Logic* 17, 1 (1988), 1–26. http://www.jstor.org/stable/30226385

[10] Ronald Fagin and Joseph Y Halpern. 1987. Belief, awareness, and limited reasoning. *Artificial intelligence* 34, 1 (1987), 39–76.

[11] Claudia Fernández-Fernández. 2019. Awareness Logic: an Epistemological Defence Correlations between Awareness Logic and Epistemology. *Kairos. Journal of Philosophy & Science* 22, 1 (2019), 72–85.

[12] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14 (2004), 349–379.

[13] Valentin Goranko. 2001. Coalition games and alternating temporal logics. In *Theoretical Aspects Of Rationality And Knowledge: Proceedings of the 8 th conference on Theoretical aspects of rationality and knowledge*, Vol. 8. 259–272.

[14] Google Nest Help. 2024. Cold weather battery charging behavior in Nest cameras and doorbells. [Online; accessed 09-Jan-2024].

[15] José Hernández-Orallo and David L Dowe. 2013. On potential cognitive abilities in the machine kingdom. *Minds and Machines* 23 (2013), 179–210.

[16] Andreas Herzig and Nicolas Troquard. 2006. Knowing how to play: uniform choices in logics of agency. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems* (Hakodate, Japan) *(AAMAS '06)*. Association for Computing Machinery, New York, NY, USA, 209–216. https://doi.org/10.1145/1160633.1160666

[17] John F. Horty. 2001. *Agency and Deontic Logic.* Oxford University Press, New York, NY. https://doi.org/10.1093/0195134613.001.0001

[18] John F Horty and Nuel Belnap. 1995. The deliberative STIT: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic* 24, 6 (1995), 583–644.

[19] Atoosa Kasirzadeh and Victoria McGeer. 2023. Intelligent Capacities in Artificial Systems. In *Artificial Dispositions: Investigating Ethical and Metaphysical Issues*, William A. Bauer and Anna Marmodoro (Eds.). Bloomsbury, London, UK.

[20] Jason Konek. 2016. Probabilistic Knowledge and Cognitive Ability. *Philosophical Review* 125, 4 (2016), 509–587. https://doi.org/10.1215/00318108-3624754

[21] Angelika Kratzer. 1977. What 'Must' and 'Can' Must and Can Mean. *Linguistics and Philosophy* 1, 3 (1977), 337–355. https://doi.org/10.1007/bf00353453

[22] Angelika Kratzer. 1979. Conditional necessity and possibility. In *Semantics from different points of view*. Springer, New York, NY, 117–147.

[23] Angelika Kratzer. 1981. *The Notional Category of Modality.* De Gruyter, Berlin, Boston, 38–74. https://doi.org/10.1515/9783110842524-004

[24] Sarit Kraus and Daniel Lehmann. 1988. Knowledge, belief and time. *Theoretical Computer Science* 58, 1-3 (1988), 155–174.

[25] Emiliano Lorini. 2013. Temporal logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics* 23, 4 (2013), 372–399.

[26] Emiliano Lorini and Pengfei Song. 2023. A computationally grounded logic of awareness. *Journal of Logic and Computation* 33, 6 (2023), 1463–1496.

[27] Tesla Model Y Owner's manual. 2023. *Traffic Light and Stop Sign Control.* TESLA, INC. https://www.tesla.com/ownersmanual/modely/en_us/GUID-A701F7DC-875C-4491-BC84-605A77EA152C.html

[28] Matt McFarland. 2020. *Uber self-driving car operator charged in pedestrian death.* CNN Business. https://www.cnn.com/2020/09/18/cars/uber-vasquez-charged/index.html

[29] Matt McFarland. 2022. *Tesla Autopilot's safety questioned after latest fatal motorcycle crash.* CNN Business. https://www.cnn.com/2022/10/17/business/tesla-motorcycle-crashes-autopilot/index.html

[30] Alfred R. Mele. 2003. Agents' Abilities. *Noûs* 37, 3 (2003), 447–470. https://doi.org/10.1111/1468-0068.00446

[31] Elliott Mendelson. 2009. *Introduction to mathematical logic.* CRC press, Boca Raton, FL.

[32] Robert C Moore et al. 1985. *A formal theory of knowledge and action.* Vol. 31. Center for the Study of Language and Information, Stanford, CA.

[33] Pavel Naumov and Jia Tao. 2018. Together we know how to achieve: An epistemic logic of know-how. *Artificial Intelligence* 262 (2018), 279–300.

[34] Pavel Naumov and Jia Tao. 2020. An epistemic logic of blameworthiness. *Artificial Intelligence* 283 (2020), 103269.

[35] Duncan Pritchard. 2010. Cognitive ability and the extended cognition thesis. *Synthese* 175, Suppl 1 (2010), 133–151.

[36] Thomas Reid. 1819. *Essays on the active powers of the human mind.* Vol. 3. Bell, Edinburgh.

[37] Catherine M. Robb. 2020. Talent Dispositionalism. *Synthese* 198, 9 (2020), 8085–8102. https://doi.org/10.1007/s11229-020-02559-6

[38] Gilbert Ryle. 1949. *The Concept of Mind: 60Th Anniversary Edition.* Hutchinson & Co, New York.

[39] Carolyn Said. 2018. *Video shows Uber robot car in fatal accident did not try to avoid woman.* SFGATE. https://www.sfgate.com/business/article/Uber-video-shows-robot-car-in-fatal-accident-did-12771938.php

[40] A. Carlisle Scott, William J. Clancey, Randall Davis, and Edward H. Shortliffe. 1977. Explanation Capabilities of Production-Based Consultation Systems. *American Journal of Computational Linguistics* (Feb. 1977), 1–50. https://aclanthology.org/J77-1006 Microfiche 62.

[41] Faiz Siddiqui and Jeremy B. Merrill. 2023. *17 fatalities, 736 crashes: The shocking toll of Tesla's Autopilot.* Washington Post. https://www.washingtonpost.com/technology/2023/06/10/tesla-autopilot-crashes-elon-musk/

[42] Jack Spencer. 2017. Able to Do the Impossible. *Mind* 126, 502 (2017), 466–497. https://doi.org/10.1093/mind/fzv183

[43] Jason Stanley and Timothy Williamson. 2001. Knowing How. *Journal of Philosophy* 98, 8 (2001), 411–444.

[44] William R Swartout. 1985. Explaining and justifying expert consulting programs. In *Computer-assisted medical decision making*. Springer, New York, NY, 254–271.

[45] Johan Van Benthem. 2001. Games in Dynamic-Epistemic Logic. *Bulletin of Economic Research* 53, 4 (2001), 219–248.

[46] Peter Van Inwagen. 1983. *An essay on free will.* Oxford University Press, New York.

[47] Ming Xu. 1998. Axioms for deliberative stit. *Journal of Philosophical Logic* 27 (1998), 505–552.

APPENDIX

Lemma 4.4. $\vdash \mathsf{K}_a\varphi \to \mathsf{K}_a\mathsf{K}_a\varphi$.

Proof. Formula $\mathsf{K}_a\neg\mathsf{K}_a\varphi \to \neg\mathsf{K}_a\varphi$ is an instance of the Truth axiom. Thus, $\vdash \mathsf{K}_a\varphi \to \neg\mathsf{K}_a\neg\mathsf{K}_a\varphi$ by contraposition. Hence, considering the following instance of the Negative Introspection axiom: $\neg\mathsf{K}_a\neg\mathsf{K}_a\varphi \to \mathsf{K}_a\neg\mathsf{K}_a\neg\mathsf{K}_a\varphi$, we have

$$\vdash \mathsf{K}_a\varphi \to \mathsf{K}_a\neg\mathsf{K}_a\neg\mathsf{K}_a\varphi. \tag{11}$$

At the same time, $\neg\mathsf{K}_a\varphi \to \mathsf{K}_a\neg\mathsf{K}_a\varphi$ is an instance of the Negative Introspection axiom. Thus, $\vdash \neg\mathsf{K}_a\neg\mathsf{K}_a\varphi \to \mathsf{K}_a\varphi$ by the law of contrapositive in the propositional logic. Hence, by the Necessitation inference rule, $\vdash \mathsf{K}_a(\neg\mathsf{K}_a\neg\mathsf{K}_a\varphi \to \mathsf{K}_a\varphi)$. Thus, by the Distributivity axiom and the Modus Ponens inference rule, $\vdash \mathsf{K}_a\neg\mathsf{K}_a\neg\mathsf{K}_a\varphi \to \mathsf{K}_a\mathsf{K}_a\varphi$. The latter, together with statement (11), implies the statement of the lemma by propositional reasoning. □

Lemma 4.5. If $X, \varphi \vdash \psi$, then $X \vdash \varphi \to \psi$.

Proof. Suppose that sequence $\psi_1, \dots, \psi_n$ is a proof from set $X \cup \{\varphi\}$ and the theorems of our logical system that uses the Modus Ponens inference rule only. In other words, for each $k \leq n$, either

(1) $\vdash \psi_k$, or
(2) $\psi_k \in X$, or
(3) ψ_k is equal to φ, or
(4) there are $i, j < k$ such that formula ψ_j is equal to $\psi_i \to \psi_k$.

It suffices to show that $X \vdash \varphi \to \psi_k$ for each $k \leq n$. We prove this by induction on k through considering the four cases above separately.

Case 1: $\vdash \psi_k$. Note that $\psi_k \to (\varphi \to \psi_k)$ is a propositional tautology, and thus, is an axiom of our logical system. Hence, $\vdash \varphi \to \psi_k$ by the Modus Ponens inference rule. Therefore, $X \vdash \varphi \to \psi_k$.
Case 2: $\psi_k \in X$. Note again that $\psi_k \to (\varphi \to \psi_k)$ is a propositional tautology, and thus, is an axiom of our logical system. Therefore, by the Modus Ponens inference rule, $X \vdash \varphi \to \psi_k$.
Case 3: formula ψ_k is equal to φ. Thus, $\varphi \to \psi_k$ is a propositional tautology. Therefore, $X \vdash \varphi \to \psi_k$.
Case 4: formula ψ_j is equal to $\psi_i \to \psi_k$ for some $i, j < k$. Thus, by the induction hypothesis, $X \vdash \varphi \to \psi_i$ and $X \vdash \varphi \to (\psi_i \to \psi_k)$. Note that formula $(\varphi \to \psi_i) \to ((\varphi \to (\psi_i \to \psi_k)) \to (\varphi \to \psi_k))$ is a propositional tautology. Therefore, $X \vdash \varphi \to \psi_k$ by applying the Modus Ponens inference rule twice. □

Lemma 4.6. If $\varphi_1 \wedge \dots \wedge \varphi_n \vdash \psi$, then $\mathsf{K}_a\varphi_1, \dots, \mathsf{K}_a\varphi_n \vdash \mathsf{K}_a\psi$.

Proof. Assumption $\varphi_1 \wedge \dots \wedge \varphi_n \vdash \psi$, by Lemma 4.5 applied n times, implies that $\vdash \varphi_1 \to (\varphi_2 \to \dots (\varphi_n \to \psi) \dots)$. Thus, by the Necessitation inference rule,

$$\vdash \mathsf{K}_a(\varphi_1 \to (\varphi_2 \to \dots (\varphi_n \to \psi) \dots)).$$

Hence, by the K-Distributivity axiom and the Modus Ponens rule,

$$\vdash \mathsf{K}_a\varphi_1 \to \mathsf{K}_a(\varphi_2 \to \dots (\varphi_n \to \psi) \dots).$$

Then, again by the Modus Ponens rule,

$$\mathsf{K}_a\varphi_1 \vdash \mathsf{K}_a(\varphi_2 \to \dots (\varphi_n \to \psi) \dots).$$

Therefore, $\mathsf{K}_a\varphi_1, \dots, \mathsf{K}_a\varphi_n \vdash \mathsf{K}_a\psi$ by applying the previous steps $(n-1)$ more times. □