Reasoning about Ignorance and Beliefs

Alessandro Aldini, Pierluigi Graziani, and Mirko Tagliaferri

University of Urbino Carlo Bo

Abstract. When building artificial agents that have to make decisions, understanding what follows from what they know or believe is mandatory, but it is also important to understand what happens when those agents ignore some facts. This becomes especially relevant when such agents ignore their ignorance, since this hinders their ability of seeking the information they are missing. Given this fact, it might prove useful to clarify in which circumstances ignorance is present and what might cause an agent to ignore that he/she is ignoring. This paper is an attempt at exploring those facts. In the paper, the relationship between ignorance and beliefs is analysed. In particular, three doxastic effects are discussed, showing that they can be seen as a cause of ignorance. The effects are formalized in a bi-modal formal language for knowledge and belief and it is shown how ignorance follows directly from those effects. Moreover, it is shown that negative introspection is the culprit of the passage between simply ignoring a fact and ignoring someone's ignorance about that fact. Those results could prove useful when artificial agents are designed, since modellers would be aware of which conditions are mandatory to avoid deep forms of ignorance; this means that those artificial agents would be able to infer which information they are ignoring and they could employ this fact to seek it and fill the gaps in their knowledge/belief base.

Keywords: Ignorance · Beliefs · Modal Logics.

1 Introduction

A sub-field of artificial intelligence is the one that concentrates on building *expert systems*. An expert system (ES) is a computer system that tries to emulate the decision-making abilities of human beings. ESs often rely on knowledge bases, which can be employed by the systems to infer new information and, thus, allow for better decisions [21]. In this respect, modal logic provides an invaluable contribution to the modelling of such systems, since formal systems of epistemic logic can satisfactorily represent knowledge and inferences based on knowledge [17, 3]. Those modal languages are even more impactful when finer grained cognitive phenomena are taken into consideration. For instance, BDI (Belief-Desire-Intention) intelligent systems can decide which plans are better to perform and then perform them, thus increasing their efficiency compared to expert systems that rely only on knowledge bases [16, 23, 20, 8, 22]. This shall not come as a surprise; the more aspects of human cognitive infrastructures systems can emulate, the more the actions and plans of those systems will resemble those of humans[1].

[1] See [4, 18, 1] for some recent and interesting applications of BDI systems.

While contemporary systems have become extremely good at emulating positive cognitive elements of human decision-making, one thing which is neglected is the impact of ignorance.[2] Seldom systems make explicit reference to ignorance, despite the abundance of formal work on the notion[13, 9, 12, 19, 7]. This paper is an attempt at showing that even when ignorance is not modelled directly into intelligent systems, if those systems have representations of beliefs and certain doxastic effects take place, then, it can be claimed that the systems are ignoring. In order to achieve this goal, a thorough investigation of ignorance within a logical framework is provided. Having a clear idea of how ignorance might be modelled using modal logic and which are the relationships with other cognitive phenomena such as knowledge and belief can greatly enhance the deductive powers of intelligent systems employing those improved formal languages to make their inferences. Moreover, once the formal framework is clear, it is possible to reason about higher-orders of ignorance, to allow intelligent systems to understand which dangerous cognitive stance they should avoid in order to not fall within the black-hole of ignoring to ignore. All of this will be obtained assuming a straightforward definition of ignorance, that can be assumed to be present even when not explicitly modelled into intelligent systems. The aim of the paper is thus to show how doxastic[3] phenomena relate to ignorance and, furthermore, to show what principles must be implemented into an intelligent system to avoid higher-order instances of ignorance.

The structure of the paper will be the following: in section 2, an introduction to the logic of ignorance and beliefs is provided. In section 3, the relation between believing and ignoring is explored and explained, providing insights and novel results into possible ways ignorance can emerge. In section 4, previously proven results [7] on the relationship between higher-orders[4] of ignorance are discussed. In section 5, it will be shown how the lack of negative introspection can be seen as a major cause of second-order ignorance (i.e., ignorance of ignorance), providing novel insights into the relationship between first-order ignorance and second-order ignorance. Finally, concluding remarks and possible ventures for the future will follow.

2 Logic for ignorance and beliefs

The origin of the formal discussion on ignorance can be attributed to Jaakko Hintikka's seminal work *Knowledge and Belief: An Introduction to the Logic of the Two Notions* [11]. In his book, Hintikka provides a propositional axiomatization of the two notions of knowledge and belief, providing insights also on other

[2] It is necessary to clarify that the term *ignorance* employed in this paper is given a specific meaning, i.e., to *not know/not be aware of* something.

[3] In this paper, the term doxastic will always refer to the act of believing.

[4] In the paper, ignorance will always be indicated with a specific order, which indicates the depth of the ignoring phenomenon. First-order ignorance means that a given fact is ignored; second-order ignorance means that it is ignored that a given fact is ignored, and so forth.

cognitive notions such as that of ignorance. This work is important because it is the first attempt to try to axiomatize ignorance employing an axiomatization of knowledge as a starting point. Following this path, Hintikka gave birth to the classical approach of formally defining ignorance in terms of lack of knowledge. Specifically, for Hintikka, ignoring a specific proposition ϕ is equivalent to not knowing whether ϕ is true or false (formally $I(\phi) =_{def} \neg K(\phi) \wedge \neg K(\neg\phi)$). This definition of ignorance, while natural, might be considered stronger than the common notion of ignorance interpreted as not knowing, since people tend to use the phrase "ignoring ϕ" to simply stand for "not knowing that ϕ" (formally $I(\phi) =_{def} \neg K(\phi)$). While this might be true, when dealing with artificial agents, the cognitive stance of those agents can be interpreted as the stronger one, therefore it makes sense to follow the classical approach and employ Hintikka's original definition.

2.1 Defining the formal framework

In this paper, the formal definition of ignorance of Hintikka [11] will be assumed. Moreover, the two notions of knowledge and belief that will be discussed are going to be interpreted in the language of propositional modal logic, using, as a semantic basis, Kripke structures. In particular, a bimodal language $\mathcal{L}$ will be employed. For brevity purposes, the syntactic definition of the language will be provided, but the semantics will not be presented[5].

Definition 1 (Logical Language for Knowledge and Belief). *Given a countable set At of primitive propositions $p_1, \ldots, p_n$ the bimodal logical language $\mathcal{L}$ is defined by the set of all formulas obtained through the following recursive procedure:*

$\phi := p_i \mid \neg\phi \mid \phi \wedge \phi \mid K(\phi) \mid B(\phi)$ *with* $p_i \in At$

All the other Boolean connectives are defined in the standard way. The main modalities of the language will be K and B, where $K(\phi)$ should be read as "ϕ is known" and will be called *knowledge formula*; $B(\phi)$ should be read as "ϕ is believed" and will be called *belief formula*. Finally, ignorance (I) is defined in terms of knowledge in the following way: $I(\phi) =_{def} \neg K(\phi) \wedge \neg K(\neg\phi)$, where $I(\phi)$ will be called an *ignorance formula*[6].

In the language $\mathcal{L}$ different readings for the ignorance formulas will be employed. It will be said that a formula ϕ is *first-order ignored*, whenever there is only one ignorance operator applied to it, the simplest case being $I(\phi)$. It will be said that a formula ϕ is *second-order ignored*, whenever there are at least, and no more than, two nested ignorance operators applied to ϕ. Again, the simplest case is $I(I(\phi))$. Higher-order instances of ignorance follow a similar path.

[5] The interested reader is referred to [17] for a standard presentation of Kripke structures.

[6] Note that an ignorance formula could represent instances of ignorance of any order, depending on how many occurrences of ignorance operators appear in the formula ϕ.

Some particular properties of the two notions of knowledge and beliefs will be assumed. For the notion of knowledge, it will be assumed that knowledge is *factual* (T) and *positively introspective* (4). The factuality (or truthfulness) of knowledge is pretty straightforward: this comes mainly from philosophical reflections on the notion of knowledge, which is taken to be a rigorous cognitive phenomenon strongly tied with truth, i.e., only true things might be known. In fact, the strength of this axiom is what distinguishes proper knowledge from simple beliefs. Beliefs might be false, but knowledge never is. The positive introspective axiom might be slightly more problematic and it comes from the assumption that agents have a privileged access to their cognitive states. This might not always make sense for human beings, who can often forget what they know and thus, are unable to keep track of everything they know, but it is a reasonable assumption for expert and intelligent systems, which always explicitly compute what they know and thus have records of all the things they know, without major issues on the memory side of things. For the notion of belief, it will be assumed that *beliefs are consistent* (D). Consistency of beliefs means that someone cannot believe that a fact is both true and false at the same time. As with positive introspection, this assumption might not always be valid for ordinary human beings, especially the irrational ones; however, since expert and intelligent systems should resemble the behaviour of rational agents, having such a consistency imposition is mandatory[7]. Finally, it will be assumed that the two notions interact in the following way: knowledge will always imply belief (Int_1) and whenever something is believed, it is known that it is believed (Int_2). The first interaction axiom is commonly derived directly from the analysis of knowledge given in Plato's Theatetus [15]: in such an analysis, knowledge is taken to be justified true belief. Unfortunately, the justification component is often neglected in formal languages, even though some attempts have been made to insert it[8]; the truth component is formalized through axiom T, while the belief component is given exactly by the interaction axiom Int_1. Int_2, on the other hand, is justified using arguments similar to the ones employed for positive introspection. In fact, it is assumed that agents not only have privileged access to their knowledge, but also to their beliefs. Again, this makes perfect sense when computational systems are involved, since they often can keep a record of what they know and/or believe. Formally, all those properties are axiomatized through the following formulas:

Definition 2 (Properties of Knowledge and Belief). *The following formulas are (assumed to be) valid for knowledge:*

- ***K**: $K(\phi \to \psi) \to (K(\phi) \to K(\psi))$.*
- ***T**: $K(\phi) \to \phi$.*
- ***4**: $K(\phi) \to K(K(\phi))$.*

The following formulas are (assumed to be) valid for belief:

[7] Note that, given the semantic framework employed to interpret the two notions, those notions also distribute over implications.

[8] See, e.g., [2].

- ***B**: $B(\phi \to \psi) \to (B(\phi) \to B(\psi))$.*
- ***D**: $\neg(B(\phi) \wedge B(\neg\phi))$.*

The following formulas are assumed to be valid for the interaction between knowledge and beliefs:

- **$\mathbf{Int_1}$***: $K(\phi) \to B(\phi)$.*
- **$\mathbf{Int_2}$***: $B(\phi) \to K(B(\phi))$.*

A further axiom which will be employed in later sections of the paper, but will not be assumed in the language is the axiom of *negative introspection*, often known as axiom 5 of epistemic logic. Negative introspection is similar in spirit to positive introspection: both axioms attribute to the agents a form of transparency towards their cognition. As was said above, positive introspection allows an agent to know everything he/she knows; on the other hand, negative introspection says that an agent will always know what he/she does not know, i.e., $\neg K(\phi) \to K(\neg K(\phi))$. This axiom, while often assumed in epistemic languages employed in computer science [10], might be too demanding for artificial agents, since it would imply that those agents are aware of all the facts they do not know. Making the reasonable assumption that there are an infinite amount of unknown facts, this would mean that the artificial agent has an infinite memory to stock all those facts that it knows not to know. It will be shown later that negative introspection alone is sufficient to prevent the occurrence of higher-order instances of ignorance. Moreover, it will be shown that when negative introspection is assumed missing, then first-order ignorance and second-order ignorance are tightly tied together.

Note that in the proofs that are given in this paper, various inference rules will be employed. All those rules are standard rules of modal logic. Since indicating all the rules employed would occupy way too much space, the reader is invited to check [14] and [5] for references on all the rules that will be employed in this paper[9].

Now that all the formal details have been given, it is possible to move on to the reflections concerning the interplay between beliefs and ignorance.

3 Misbelieving, being agnostic or doubting

Understanding the origin of ignorance is quite complicated. Sometimes, it is easy to recognize if someone is ignorant about something, but it is not clear what brought about and fed this ignorance. The main issue is that ignorance is a *negative fact*, i.e., it is a lack of knowledge, and, therefore, there is no specific

[9] The abbreviations that will be employed in the proofs of this paper will all be reported here. *Ass.* will stand for “assumption”; *P. Taut.* will stand for “propositional tautology”; *Elim.* will stand for “elimination rule”; *Intr.* will stand for “introduction rule”; *Contrap.* will stand for “contrapposition”; *MP* will stand for “Modus Ponens”; *DM* will stand for “DeMorgan rules”; *DS* will stand for “disjunctive syllogism”; *Nec.* will stand for “necessitation rule”; finally *Distr.* will stand for “distributivity rule”.

moment in time when ignorance is generated; it is there the whole time, until it disappears. Simply put, ignorance is not something that can be gained, but only lost. Not having a specific moment in time during which ignorance originates makes it difficult for researchers to focus on specific acts or behaviours that can aid their understanding of the phenomenon. For this reason, a formal research on the notion of ignorance might help to understand what are the constituents of such notion and thus which other phenomena are responsible for its emergence and/or existence. Once the formal links between doxastic effects and ignorance are understood and recognized, modellers will be able to design artificial agents that are better suited to deal with ignorance and the effects ignorance has in planning and pursuing a specific goal. Specifically, three different, alternative doxastic effects will be explored, showing that those individually imply ignorance and, conversely, they are implied by ignorance, making them equivalent to ignorance. The first of those states will be called *the misbelieving effect*, the second will be called *the agnostic effect*, and, finally, the third one will be called *the doubting effect*.

Intuitively, we say that an agent is subject to the *misbelieving effect* either when the agent believes that a given fact is true, while it is false, or when the agent believes that a given fact is false, while it is true[10].

Definition 3 (Misbelieving Effect). *The misbelieving effect is represented by the following formula:*

$$(B(\phi) \wedge \neg\phi) \vee (B(\neg\phi) \wedge \phi) \tag{1}$$

The misbelieving effect is quite common. Everybody, even the most conscientious human being, will have some misbeliefs about the world that surrounds him/her. Science is full of examples: researchers constantly discover new facts that contradict what was previously thought to be true, thus highlighting many misbeliefs that were held by those scientists. Per se, the fact that this effect is so extensively spread does not cause many problems, since misbelieving, when taken in isolation, only implies ignorance and it is plausible that most scientists will admit to be ignorant about many things. However, if the misbelieving agent is not open to revise his/her beliefs, the misbelieving effect might cause dangerous issues, since both first-order ignorance and higher-order instances of ignorance will be produced.

Intuitively, we say that an agent is subject to the *agnostic effect* when the agent neither believes that a given fact is true nor believes that the fact is false.

Definition 4 (Agnostic Effect). *The agnostic effect is represented by the following formula:*

$$\neg B(\phi) \wedge \neg B(\neg\phi) \tag{2}$$

Again, also the agnostic effect is quite common. People not having an opinion about a specific matter are the prime candidates of agents which are subject to

[10] See [6] for a discussion about different aspects that relate misbelieving and ignoring.

this effect. Since they do not have an opinion about a given fact, they simply do not believe neither in the truth of the fact nor in its falsity. Note that this does not mean that they do not believe that the fact is either true or false (which is indeed a tautology and must be believed due to the necessitation rule and Int_1), but they cannot make up their mind in one direction or the other and, thus, suspend their judgement. It is not surprising that the agnostic effect causes ignorance, since the lack of beliefs is just the first step to the lack of knowledge. Again, this is not a problem if the agnostic effect is due to a suspension of judgement about the truth of a specific fact, since this is just a clear acknowledgement that such fact is ignored. The problem begins when the agnostic effect is coupled with unawareness of the possibility of believing that the fact is either true or false. As with misbelieving, also in this latter case, being agnostic does not just cause first-order ignorance, but also higher-order instances of ignorance.

Intuitively, we say that an agent is subject to the *doubting effect* when the agent believes in something which in fact holds, but he/she does not have the guarantee that such fact actually holds.

Definition 5 (Doubting Effect). *The doubting effect is represented by the following formula:*

$$(B(\phi) \land \phi \land \neg K(\phi)) \lor (B(\neg\phi) \land \neg\phi \land \neg K(\neg\phi)) \tag{3}$$

The doubting effect is similar in spirit to the misbelieving effect. Since in both cases agents do not have access to the state of the world, from a first-person perspective, it is impossible, for the agent, to recognize whether he/she is misbelieving or is simply doubtful. The main difference between the two cases is that, in the doubting effect, the lack of knowledge of the agent is explicitly specified. This specification is fundamental, since it is the main culprit of the emergence of ignorance. This does not seem to be a great surprise, since, per se, believing something that actually holds should not cause problems.

Even though it seems quite reasonable that the misbelieving effect, the agnostic effect and the doubting effect imply first-order ignorance, such facts must be proven. The existence of these proofs in standard formal systems can be given both a normative and a descriptive reading. On the normative side, they show that the intuitions about misbelieving, being agnostic and doubting are indeed well-guided and, thus, strengthen the relation between beliefs and knowledge; on the descriptive side, if it is assumed that the intuitions are justified, the proofs provided here show that classical epistemic and doxastic formal systems are well-structured and manage to properly describe real world phenomena. The proofs provided will show that the three effects presented above are individually sufficient for ignorance. Subsequently, it will also be shown that ignorance will always imply at least one of the three effects.

3.1 From misbelieving to ignoring

The conditional connecting the misbelieving effect and first-order ignorance is valid in the language $\mathcal{L}$.

Proposition 1. $((B(\phi) \wedge \neg\phi) \vee (B(\neg\phi) \wedge \phi)) \rightarrow (\neg K(\phi) \wedge \neg K(\neg\phi))$ *is valid.*

Prop. 1 says that, in $\mathcal{L}$, misbelieving and ignorance are tied together. Interestingly, this connection holds also in weaker languages, since axiom 4 and axiom B of $\mathcal{L}$ are not needed in the proof of the proposition[11].

3.2 From being agnostic to ignoring

The conditional connecting the agnostic effect and ignorance is valid in the language $\mathcal{L}$.

Proposition 2. $(\neg B(\phi) \wedge \neg B(\neg\phi)) \rightarrow (\neg K(\phi) \wedge \neg K(\neg\phi))$ *is valid.*

Prop. 2 says that, in $\mathcal{L}$, being agnostic and ignorance are tied together. Again, this result holds also in weaker languages (in fact, systems even weaker than the ones that satisfy Prop. 1), since only the interaction axiom *Int1* is needed to obtain the proof[12].

3.3 From doubting to ignoring

The conditional connecting the doubting effect and ignorance is valid in the language $\mathcal{L}$.

Proposition 3. $((B(\phi) \wedge \phi \wedge \neg K(\phi)) \vee (B(\neg\phi) \wedge \neg\phi \wedge \neg K(\neg\phi))) \rightarrow (\neg K(\phi) \wedge \neg K(\neg\phi))$ *is valid.*

Prop. 3 shows that, in $\mathcal{L}$, doubting and ignorance are tied together. Note that this connection also holds in weaker systems, since only axiom T of $\mathcal{L}$ has been used in the proof.

3.4 From ignoring to the three effects

The fact that ignorance must imply one among the three effects will now be proven.

Theorem 1. *The conditional connecting first-order ignorance to the disjunction of the three doxastic effects is valid in the language $\mathcal{L}$. Formally, this is equivalent to stating that:*

$$\begin{array}{lll}
\neg K(\phi) \wedge \neg K(\neg\phi) \rightarrow & & \\
& B(\phi) \wedge \neg\phi \text{ (Misb.)} & \vee \\
& B(\neg\phi) \wedge \phi \text{ (Misb.)} & \vee \\
& \neg B(\phi) \wedge \neg B(\neg\phi) \text{ (Agnos.)} & \vee \\
& B(\phi) \wedge \phi \wedge \neg K(\phi) \text{ (Doubt.)} & \vee \\
& B(\neg\phi) \wedge \neg\phi \wedge \neg K(\neg\phi) \text{ (Doubt.)} &
\end{array}$$

is valid.

[11] See appendix A.

[12] See appendix A.

Proof (Theorem 1).

The proof will be given by contradiction, showing that first-order ignorance is incompatible with the negation of all the three doxastic effects.

Assume:

$$\begin{array}{ll}
\neg K(\phi) \wedge \neg K(\neg\phi) & \wedge \\
\neg((B(\phi) \wedge \neg\phi) \vee (B(\neg\phi) \wedge \phi)) & \wedge \\
\neg(\neg B(\phi) \wedge \neg B(\neg\phi)) & \wedge \\
\neg((B(\phi) \wedge \phi \wedge \neg K(\phi)) \vee (B(\neg\phi) \wedge \neg\phi \wedge \neg K(\neg\phi))) &
\end{array}$$

The previous formula can be transformed into *Conjunctive Normal Form* (CNF), i.e., a series of clauses connected by $\wedge$s where each clause only contains $\vee$s. The first step to do so is to apply DeMorgan to the three clauses (double negations will also be eliminated directly):

$$\begin{array}{ll}
\neg K(\phi) \wedge \neg K(\neg\phi) & \wedge \\
\neg(B(\phi) \wedge \neg\phi) \wedge \neg(B(\neg\phi) \wedge \phi) & \wedge \\
B(\phi) \vee B(\neg\phi) & \wedge \\
\neg(B(\phi) \wedge \phi \wedge \neg K(\phi)) \wedge \neg(B(\neg\phi) \wedge \neg\phi \wedge \neg K(\neg\phi)) &
\end{array}$$

A second iteration of DeMorgan is possible (again, double negations will be eliminated directly):

$$\begin{array}{ll}
\neg K(\phi) \wedge \neg K(\neg\phi) & \wedge \\
(\neg B(\phi) \vee \phi) \wedge (\neg B(\neg\phi) \vee \neg\phi) & \wedge \\
(B(\phi) \vee B(\neg\phi)) & \wedge \\
(\neg B(\phi) \vee \neg\phi \vee K(\phi)) \wedge (\neg B(\neg\phi) \vee \phi \vee K(\neg\phi)) &
\end{array}$$

Now, a row will be given to each clause of the above CNF formula:

$$\begin{array}{lll}
\text{(a)} & \neg K(\phi) & \wedge \\
\text{(b)} & \neg K(\neg\phi) & \wedge \\
\text{(c)} & \neg B(\phi) \vee \phi & \wedge \\
\text{(d)} & \neg B(\neg\phi) \vee \neg\phi & \wedge \\
\text{(e)} & B(\phi) \vee B(\neg\phi) & \wedge \\
\text{(f)} & \neg B(\phi) \vee \neg\phi \vee K(\phi) & \wedge \\
\text{(g)} & \neg B(\neg\phi) \vee \phi \vee K(\neg\phi) &
\end{array}$$

Taking the list above as a reference, it is possible to prove that the set of formulas (a)-(g) leads to a contradiction.

Note first that clause (e) produces two separate cases, i.e., either $B(\phi)$ holds or $B(\neg\phi)$ holds. It will be shown that both cases lead to a contradiction.

Case 1:

(1)	$B(\phi)$	Ass.
(2)	$\neg B(\phi) \lor \phi$	Clause (c)
(3)	ϕ	*DS* (1)-(2)
(4)	$\neg K(\phi)$	Clause (a)
(5)	$B(\phi) \land \phi \land \neg K(\phi)$	$\land$ Intr. (1)-(3)-(4)
(6)	$\neg(\neg B(\phi) \lor \neg\phi \lor K(\phi))$	*DM* (5)
(7)	$\neg B(\phi) \lor \neg\phi \lor K(\phi)$	Clause (f)
(8)	Contradiction	(6)-(7)

Case 2:

(1)	$B(\neg\phi)$	Ass.
(2)	$\neg B(\neg\phi) \lor \neg\phi$	Clause (d)
(3)	$\neg\phi$	*DS* (1)-(2)
(4)	$\neg K(\neg\phi)$	Clause (b)
(5)	$B(\neg\phi) \land \neg\phi \land \neg K(\neg\phi)$	$\land$ Intr. (1)-(3)-(4)
(6)	$\neg(\neg B(\neg\phi) \lor \phi \lor K(\neg\phi))$	*DM* (5)
(7)	$\neg B(\neg\phi) \lor \phi \lor K(\neg\phi)$	Clause (g)
(8)	Contradiction	(6)-(7)

All cases lead to a contradiction. Therefore, at least one doxastic effect must be true whenever first-order ignorance is present. □

4 Hierarchies of Ignorance

When dealing with hierarchies of ignorance, there are at least two important aspects which require analysis. The first aspect is the one that describes the relation between first-order ignorance and second-order ignorance; the second aspect is the one that describes the relation between second-order ignorance and higher-order levels of ignorance. The importance of those aspects is based on one fundamental fact: first-order ignorance is a common phenomenon of every day life; people are ignorant about many facts and information about the world they live in. Not only common people, but also scientists and curious persons fall victim to the phenomenon of ignoring. It is an indissoluble trait of all human beings. Nonetheless, first-order ignorance is not problematic on its-own; quite the opposite, first-order ignorance is what often stimulates the genuine curiosity that pushes human beings towards making new discoveries and increasing their overall knowledge. What can be considered problematic is the ignorance of ignorance (second-order ignorance), since this phenomenon precludes the possibility of dissipating first-order ignorance, given that people do not have the stimulus to understand something they are not even aware of being ignorant about. This should highlight the importance of understanding and exploring what is the relation between first-order ignorance and second-order ignorance. Once the interplay between the two phenomena is clear, it is possible to design strategies that lock the passage from the former to the latter.

The second aspect (the relation between second-order ignorance and higher-orders of ignorance) is important for similar reasons. Once it is admitted that some forms of second-order ignorance are unavoidable, it might be good to know that such second-order ignorance exists, i.e., to know that one is second-order ignorant about something. At least, such knowledge would stimulate persons to work on their ignorance, in order to avoid it.

While the first aspect is still obscure in the literature on the formal representation of ignorance and will be explored in the next section of this paper, the second aspect has been well explored by Kit Fine in his paper "Ignorance of ignorance" [7][13]. In his paper, Fine shows that second-order ignorance and higher-orders of ignorance are tightly tied together. Once second-order ignorance is present, an agent is doomed to the black hole of higher-order levels of ignorance.

Those aspects about the hierarchies of ignorance are especially important when strategies for modelling artificial agents are taken into consideration. This is due to the fact that if modellers do not pay enough attention, those artificial agents might end up falling victims of second-order ignorance and, subsequently, to higher-orders of ignorance; they would thus be unable to recognize that they are missing some information and would not look for it.

Theorem 2 (Fine's Ignorance Theorem). *Second-order ignorance implies higher-orders of ignorance. Specifically, second-order ignorance implies third-order ignorance. Third-order ignorance implies fourth-order ignorance and so forth.*

The proof of this statement is straightforward and only requires a few formal definitions and a few lemmas. First, the notion of *Rumsfeld ignorance* of ϕ is introduced. Intuitively, someone is Rumsfeld ignorant when he is first-order ignorant about ϕ and does not know it.

Definition 6 (Rumsfeld Ignorance). *Rumsfeld ignorance of a formula ϕ is represented by the formula*

$$I_R(\phi) =_{def} I(\phi) \wedge \neg K(I(\phi))$$

where $I(\phi)$ is a first-order ignorance formula.

Lemma 1 (From second-order ignorance to first-order ignorance).
Second-order ignorance implies first-order ignorance, i.e., $I(I(\phi)) \rightarrow I(\phi)$ is a valid formula of $\mathcal{L}$.[14]

[13] It should be pointed out that Fine does not use the terms "first-order ignorance", "second-order ignorance" and so on. However, to maintain coherence with the rest of the paper, those terms will be employed when the concepts expressed by Fine are aligned with the meanings attributed to those terms in this paper.

[14] Proofs of lemmas will not be provided. If the reader is interested, in [7] it is possible to find all the details concerning the lemmas which are introduced here. The only important detail is that Fine provides proofs in the axiomatic system $S4$, which is a system that defined languages weaker than the one employed in this paper. Therefore, every proof provided by Fine could be easily reproduced inside $\mathcal{L}$

Lemma 2 (From second-order ignorance to Rumsfeld ignorance). *Second-order ignorance implies Rumsfeld ignorance, i.e., $I(I(\phi)) \rightarrow I_R(\phi)$ is a valid formula of $\mathcal{L}$.*

Lemma 3 (From Rumsfeld ignorance to second-order ignorance). *Rumsfeld ignorance implies second-order ignorance, i.e., $I_R(\phi) \rightarrow I(I(\phi))$ is a valid formula of $\mathcal{L}$.*

Lemma 4. *One cannot know that he/she is Rumsfeld ignorant, i.e., $\neg K(I_R(\phi))$ is a valid formula of $\mathcal{L}$.*

A further lemma which will be useful later is the following[15].

Lemma 5. *If someone is second-order ignorant, then one does not know to be second-order ignorant. Formally, $I(I(\phi)) \rightarrow \neg K(I(I(\phi)))$.*

It is now possible to prove Fine's main result about the relationship between second-order ignorance and higher-orders of ignorance.

Proof (Theorem 2).

(1)	$I(I(\phi))$	Ass.
(2)	$I(I(\phi)) \rightarrow \neg K(I(I(\phi)))$	Lemma 5
(3)	$\neg K(I(I(\phi)))$	*MP* (1)-(2).
(4)	$K(\neg I(I(\phi))) \rightarrow \neg(I(I(\phi)))$	Axiom T.
(5)	$I(I(\phi)) \rightarrow \neg K(\neg I(I(\phi)))$	Contrap. (4).
(6)	$\neg K(\neg I(I(\phi)))$	*MP* (1)-(5).
(7)	$\neg K(I(I(\phi))) \wedge \neg K(\neg I(I(\phi)))$	$\wedge$ Intr. (3)-(6).
(8)	$I(I(I(\phi)))$	Definition of (7).

What Theorem 2 shows is that there is a deep connection between second-order ignorance and higher-order levels of ignorance. In fact, as soon as someone is second-order ignorant, there is no possibility that he/she escapes the dark hole of ignorance on his/her own. Once this is well understood, it becomes evident why deep investigations on the relation between first-order ignorance and second-order ignorance are required. Once it is established what causes second-order ignorance in the presence of first-order ignorance, it might be possible to stop agents from crossing the *event-horizon* of the black hole which is second-order ignorance. The rest of the paper will be dedicated to the exploration of such relation.

5 The Birth of Second-Order Ignorance

As it has been shown in the previous section, once an agent steps into second-order ignorance, he/she also enters the black hole of higher-order levels of ignorance, without having much hope to escape, since, formally, this black hole

[15] Fine proves such lemma while proving his main theorem. However, to make the proof easier to read, this lemma will be given separately. The proof of such lemma can be found in appendix A.

is inescapable employing the resources internal to the language. Assuming that first-order ignorance phenomena are common, it is important, when modelling artificial agents, to avoid possible passages from first-order ignorance to second-order ignorance, so that the black hole of higher-order levels of ignorance is avoided. Interestingly, negative introspection is an incredibly powerful cognitive phenomenon that can block the passage between first-order ignorance and second-order ignorance[16].

Theorem 3. *If Axiom* 5 *(negative introspection) of epistemic logic is assumed in the language $\mathcal{L}$, then $I(\phi) \to \neg I(I(\phi)))$ holds in $\mathcal{L}$.*

Proof (Theorem 3).

(1)	$I(\phi)$	Ass.
(2)	$\neg K(\phi) \land \neg K(\neg\phi)$	Definition of (1).
(3)	$\neg K(\phi)$	$\land$ Elim. (2).
(4)	$\neg K(\phi) \to K(\neg K(\phi))$	Axiom 5.
(5)	$K(\neg K(\phi))$	*MP* (3)-(4).
(6)	$\neg K(\neg\phi)$	$\land$ Elim. (2).
(7)	$\neg K(\neg\phi) \to K(\neg K(\neg\phi))$	Axiom 5.
(8)	$K(\neg K(\neg\phi))$	*MP* (6)-(7).
(9)	$K(\neg K(\phi) \land \neg K(\neg\phi))$	$\land$ Distr. (5)-(8).
(10)	$K(I(\phi))$	Definition of (9).
(11)	$K(I(\phi)) \lor K(\neg I(\phi))$	$\lor$ Intr. (10).
(12)	$\neg(\neg K(I(\phi)) \land \neg K(\neg I(\phi)))$	*DM* (11)
(13)	$\neg I(I(\phi)))$	Definition of (12).

It can therefore be safely claimed that negative introspection is an exceptionally effective measure to avoid the black hole of higher levels of ignorance. However, as discussed in section 2, assuming that artificial agents possess the deep introspection that axiom 5 requires might be too much. Unfortunately, a direct negation of negative introspection can become the main culprit in the spread of second-order ignorance. That means that even though it is reasonable to assume that agents are not negatively introspective, it is important to avoid that agents are *completely* non-negatively introspective, as this would tie together first-order ignorance and second-order ignorance, as stated by the following theorem:

Theorem 4 (From first-order ignorance to second-order ignorance). $I(\phi) \land (\neg K(\phi) \land \neg K(\neg K(\phi))) \to I(I(\phi))$ *holds in $\mathcal{L}$.*

Proof (Theorem 4).

The proof will be given by contradiction.

[16] This result is not novel to this paper, but is well known in the logical literature on formalizing ignorance

(1)	$I(\phi)$	Ass.
(2)	$\neg K(\phi) \land \neg K(\neg K(\phi))$	Ass.
(3)	$\neg I(I(\phi))$	Ass.
(4)	$\neg(\neg K(I(\phi)) \land \neg K(\neg I(\phi)))$	Def. of (3).
(5)	$K(I(\phi)) \lor K(\neg I(\phi))$	*DM* (4).
(6)	$K(I(\phi))$	Ass.
(7)	$I(\phi) \rightarrow \neg K(\phi)$	P. Taut.
(8)	$K(I(\phi) \rightarrow \neg K(\phi))$	Nec. (7).
(9)	$K(I(\phi) \rightarrow \neg K(\phi)) \rightarrow K(I(\phi) \rightarrow K(\neg K(\phi))$	Axiom K.
(10)	$K(I(\phi) \rightarrow K(\neg K(\phi))$	*MP* (8)-(9).
(11)	$K(\neg K(\phi))$	*MP* (6)-(10).
(12)	$\neg K(\neg K(\phi))$	$\land$ Elim. (2).
(13)	Contradiction	(11)-(12).
(14)	$K(\neg I(\phi))$	Ass.
(15)	$K(\neg I(\phi)) \rightarrow \neg I(\phi)$	Axiom T.
(16)	$\neg I(\phi)$	*MP* (14)-(15)
(17)	Contradiction	(1)-(16).

Since both clauses of $K(I(\phi)) \lor K(\neg I(\phi))$ lead to a contradiction, it must follow that $\neg(K(I(\phi)) \lor K(\neg I(\phi)))$, which is equivalent to $I(I(\phi))$. □

6 Conclusion and future works

In the paper, three possible conditions that make ignorance emerge have been proposed, showing that those conditions are both sufficient and necessary for ignorance to emerge. Those conditions were given in terms of beliefs and thus are employable to enrich previously proposed BDI-frameworks that model intelligent systems. This is especially important, if the modellers want to allow the intelligent system to avoid the black-hole of higher-orders of ignorance. While it might not seem a great improvement, it should be noted that a system which is unaware of being ignorant, will never be in a position to question such ignorance and, thus, will always be unable to produce plans to achieve extra information and make better decisions. In the paper, it has also been shown what can cause the passage between basic ignorance and higher-order levels of ignorance, providing insights on what should be explicitly avoided by intelligent systems. What shall be done in the future is to explore if the negative introspection condition that bridges basic and higher-order levels of ignorance can be expressed through more specific belief conditions that can be tailored to the specific cases of misbelieving, agnosticism and doubting.

References

1. Adam, C., Gaudou, B., "BDI agents in social simulations: a survey", The Knowledge Engineering Review, Vol. 31 (3), pp. 207–238, (2016).
2. Artemov, S., Fitting, M., "Justification Logic", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/sum2020/entries/logic-justification/¿.

3. Baltag, A., Renne, B., “Dynamic Epistemic Logic”, The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/win2016/entries/dynamic-epistemic/¿.
4. Caillou, P., Gaudou, B., Grignard, A., Truong, C. Q., Taillandier, P., “A Simple-to-use BDI architecture for Agent-based Modeling and Simulation”, Proceedings of the 11th Conference of the European Social Simulation Association (ESSA 2015), (2015).
5. Chellas, B., “Modal Logic: An Introduction”, Cambridge University Press, (1980).
6. Fano, V., Graziani, P., “A working hypothesis for the logic of radical ignorance”, Synthese, (2020): https://doi.org/10.1007/s11229-020-02681-5.
7. Fine, K., “Ignorance of Ignorance”, Synthese, Vol. 195 (9), pp. 4031–4045, (2018).
8. Georgeff, M., Pell, B., Pollack, M., Tambe, M., Wooldridge, M., “The Belief-Desire-Intention Model of Agency.” In: Müller J.P., Rao A.S., Singh M.P. (eds) Intelligent Agents V: Agents Theories, Architectures, and Languages. ATAL 1998, (1999).
9. Halpern, J.Y., “A Theory of Knowledge and Ignorance for Many Agents”, Journal of Logic and Computation, Vol. 7 (1), pp. 79–108, (1997).
10. Halpern, J., Moses., Y., Fagin, R., Vardi, M., “Reasoning about Knowledge”, The MIT Press, (1995).
11. Hintikka, J., “Knowledge and Beliefs: An Introduction to the Logic of the Two Notions”, Cornell University Press, (1962).
12. van der Hoek, W., Lomuscio, A. “A Logic for Ignorance.”. In: Leite J., Omicini A., Sterling L., Torroni P. (eds) Declarative Agent Languages and Technologies. DALT 2003, (2004).
13. Meyer, C., van der Hoek, W. “Epistemic Logic for AI and Computer Science”, Cambridge University Press, (1995).
14. Mints, G., “Natural Deduction for Propositional Logic.”, In: A Short Introduction to Intuitionistic Logic. The University Series in Mathematics. Springer, pp. 9–22, (2002).
15. Plato, “Theaetetus”, traslation by McDowell, J., Oxford University Press, 2014.
16. Rao, A., Georgeff, M., “Formal models and decision procedures for multi-agent systems”, Technical Note, AAII, (1995).
17. Rendsvig, R., Symons, J., “Epistemic Logic”, The Stanford Encyclopedia of Philosophy (Summer 2019 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/sum2019/entries/logic-epistemic/¿.
18. Smitha Rao, M. S, Jyothsna, A. N., “BDI: Applications and Architectures”, International Journal of Engineering Research and Technology, Vol. 2 (2), (2013).
19. Steinsvold, C., “A Note on Logics of Ignorance and Borders”, Notre Dame Journal of Formal Logic, Vol. 49 (4), pp. 385–392, (2008).
20. Souza, M., “Choices that make you change your mind : a dynamic epistemic logic approach to the semantics of BDI agent programming languages”, Ph.D. Thesis, University of Rio Grande, (2016).
21. Tripathi, K. P., “A Review on Knowledge-based Expert System: Concept and Architecture”, IJCA Special Issue on “Artificial Intelligence Techniques - Novel Approaches and Practical Applications, 2011.
22. Wooldridge, M., “Reasoning about Rational Agents”, MIT Press, (2000).
23. Wooldridge, M., “Practical reasoning with procedural knowledge.”. In: Gabbay D.M., Ohlbach H.J. (eds) Practical Reasoning. FAPR 1996, Springer, (1996).

A Formal Proofs

Proof (Proposition 1).

The proof will be split into two parts, showing that each disjunct of the antecedent of the conditional implies the consequent of the conditional.

Case 1: The proof will be given directly.

(1)	$B(\phi) \wedge \neg\phi$	Ass.
(2)	$B(\phi)$	$\wedge$ Elim. (1).
(3)	$\neg\phi$	$\wedge$ Elim. (1).
(4)	$K(\phi) \rightarrow \phi$	Axiom T.
(5)	$\neg\phi \rightarrow \neg K(\phi)$	Contrap. (4).
(6)	$\neg K(\phi)$	MP (3)-(5).
(7)	$\neg(B(\phi) \wedge B(\neg\phi))$	Axiom D.
(8)	$\neg B(\phi) \vee \neg B(\neg\phi)$	DM (7).
(9)	$\neg B(\neg\phi)$	DS (2)-(8).
(10)	$K(\neg\phi) \rightarrow B(\neg\phi)$	Axiom Int_1.
(11)	$\neg B(\neg\phi) \rightarrow \neg K(\neg\phi)$	Contrap. (10).
(12)	$\neg K(\neg\phi)$	MP (9)-(11).
(13)	$\neg K(\phi) \wedge \neg K(\neg\phi)$.	$\wedge$ Intr. (6)-(12).

Case 2: The proof will be given directly.

(1)	$B(\neg\phi) \wedge \phi$	Ass.
(2)	$B(\neg\phi)$	$\wedge$ Elim. (1).
(3)	ϕ	$\wedge$ Elim. (1).
(4)	$\neg(B(\phi) \wedge B(\neg\phi))$	Axiom D.
(5)	$\neg B(\phi) \vee \neg B(\neg\phi)$	DM (4).
(6)	$\neg B(\phi)$	DS (2)-(5).
(7)	$K(\phi) \rightarrow B(\phi)$	Axiom Int_1.
(8)	$\neg B(\phi) \rightarrow \neg K(\phi)$	Contrap. (10).
(9)	$\neg K(\phi)$	MP (6)-(8).
(10)	$K(\neg\phi) \rightarrow \neg\phi$	Axiom T.
(11)	$\phi \rightarrow \neg K(\neg\phi)$	Contrap. (10).
(12)	$\neg K(\neg\phi)$	MP (3)-(11).
(13)	$\neg K(\phi) \wedge \neg K(\neg\phi)$.	$\wedge$ Intr. (9)-(12).

Proof (Proposition 2).

The proof will be given directly.

(1)	$\neg B(\phi) \wedge \neg B(\neg\phi)$	Ass.
(2)	$\neg B(\phi)$	$\wedge$ Elim. (1).
(3)	$\neg B(\neg\phi)$	$\wedge$ Elim. (1).
(4)	$K(\phi) \to B(\phi)$	Axiom Int_1.
(5)	$\neg B(\phi) \to \neg K(\phi)$	Contrap. (4).
(6)	$\neg K(\phi)$	*MP* (2)-(5).
(7)	$K(\neg\phi) \to B(\neg\phi)$	Axiom Int_1.
(8)	$\neg B(\neg\phi) \to \neg K(\neg\phi)$	Contrap. (7).
(9)	$\neg K(\neg\phi)$	*MP* (3)-(8).
(10)	$\neg K(\phi) \wedge \neg K(\neg\phi)$.	$\wedge$ Intr. (6)-(9).

Proof (Proposition 3).

The proof will be split into two parts, showing that each disjunct of the antecedent of the conditional implies the consequent of the conditional.

Case 1: The proof will be given directly.

(1)	$B(\phi) \wedge \phi \wedge \neg K(\phi)$	Ass.
(2)	$B(\phi)$	$\wedge$ Elim. (1).
(3)	ϕ	$\wedge$ Elim. (1).
(4)	$\neg K(\phi)$	$\wedge$ Elim. (1).
(5)	$K(\neg\phi) \to \neg\phi$	Axiom T.
(6)	$\phi \to \neg K(\neg\phi)$	Contrap. (5).
(7)	$\neg K(\neg\phi)$	*MP* (3)-(6).
(8)	$\neg K(\phi) \wedge \neg K(\neg\phi)$.	$\wedge$ Intr. (4)-(7).

Case 2: The proof will be given directly.

(1)	$B(\neg\phi) \wedge \neg\phi \wedge \neg K(\neg\phi)$	Ass.
(2)	$B(\neg\phi)$	$\wedge$ Elim. (1).
(3)	$\neg\phi$	$\wedge$ Elim. (1).
(4)	$\neg K(\neg\phi)$	$\wedge$ Elim. (1).
(5)	$K(\phi) \to \phi$	Axiom T.
(6)	$\neg\phi \to \neg K(\phi)$	Contrap. (5).
(7)	$\neg K(\phi)$	*MP* (3)-(6).
(8)	$\neg K(\phi) \wedge \neg K(\neg\phi)$.	$\wedge$ Intr. (4)-(7).

Proof (Lemma 5). The proof is given by contradiction.

(1)	$I(I(\phi))$	Ass.
(2)	$I(I(\phi)) \to I_R(\phi)$	Lemma 2.
(3)	$K(I(I(\phi)))$	Ass.
(4)	$K(I(I(\phi)) \to I_R(\phi)$	*Nec.* (2).
(5)	$K(I(I(\phi)) \to I_R(\phi)) \to (K(I(I(\phi))) \to K(I_R(\phi))$	Axiom K.
(6)	$K(I(I(\phi))) \to K(I_R(\phi)$	*MP* (4)-(5).
(7)	$K(I_R(\phi)$	*MP* (3)-(6).
(8)	$\neg K(I_R(\phi))$	Lemma 4.
(9)	Contradiction	(7)-(8).

Since a contradiction has been reached, one of the assumptions must be rejected. The only assumption which can be rejected is $K(I(I(\phi)))$, thus, assuming $I(I(\phi))$, it holds that $\neg K(I(I(\phi)))$, which means that $I(I(\phi)) \rightarrow \neg K(I(I(\phi)))$ holds in $\mathcal{L}$. □