September 199 1 Report No. STAN-CS-91-1391

# A Logic for Perception and Belief

by

**Yoav Shoham and Alvaro del Val**

**Department of Computer Science**

**Stanford University**

**Stanford, California 94305**

# A Logic for Perception and Belief

Yoav Shoham
Computer Science Department
Stanford University
Stanford, CA 94305
shoham@cs.stanford.edu

Alvaro del Val
Center for the Study of Language and Information
Stanford University
Stanford, CA 94305
delval@csli.stanford.edu

September 24, 1991

Abstract

We present a modal logic for reasoning about perception and belief, captured respectively by the operators $P$ and $B$. The $B$ operator is the standard belief operator used in recent years, and the $P$ operator is similarly defined. The contribution of the paper is twofold. First, in terms of $P$ we provide a definition of ***perceptual indistinguishability,*** such as arises out of limited visual acuity. The definition is concise, intuitive (we find), and avoids traditional paradoxes. Second, we explore the bimodal $B - P$ system. We argue that the relationship between the two modalities varies among settings: The agent may or may not have confidence in its perception, may or may not be accurate in it, and so on. We therefore define a number of agent types corresponding to these various assumptions, and for each such agent type we provide a sound and complete axiomatization of the $B - P$ system.

# 1 Introduction

There is a long-standing interest in AI in defining the mental state of agents. By far the most explored components are those of ***knowledge*** and ***belief,*** versions of which have been captured in a variety of epistemic logics with well understood properties (cf. [8, 7, 9, 16]). In addition, other mental attitudes have been investigated. For example, the notion of ***commitment*** is defined in both [2] and [15]. In this paper we consider two modalities: the traditional belief modality, and a new one: ***perception.***

The reason we choose these two is that, other than being told, the main source of new beliefs is sensory input. The connection between perception and logic is difficult and multifacetted. Mackworth and Reiter [12], for example, explore the connection between vision and default reasoning. We do not address that, nor many other difficult issues in relating perception and logic. Instead, we concentrate on two issues: the notion of indistinguishability, and the relationship between perception and belief.

We define an operator ***P,*** which captures the information about the world which has been delivered by the sensors. Although on its own the operator is rather dull, it does allow us to explore the issue of perceptual indistinguishability, such as that resulting from limited visual acuity. This notion has been addressed in the philosophical logic literature (cf.[13, 5]), and at least ***once*** in AI ([3]). In this literature the main issue has been to avoid the transitivity of indistinguishability; transitivity leads to paradoxes akin to the 'heap paradox' (if two piles of sand differing in size by only one sand grain cannot be distinguished, and if indistinguishability is transitive, then no two piles can be distinguished). We provide an intuitive definition of perceptual indistinguishability in terms of ***P,*** and show that the paradox doesn't arise.

The other concern of this paper is the connection between belief and perception: we augment the language with the standard ***B*** operator for belief, and explore the connection between ***B*** and ***P.*** If the agent always believes in the truth of what it perceives, this makes things simple. This is an unacceptable assumption in general, however: There are limits to the ability of perception to register correctly the state of the world (due, e.g., to malfunctioning of the perceptual apparatus), and the agent may be aware of that. We therefore define a number of agent types, capturing different interactions between perception and belief.

The paper is organized as follows. In Section ***2*** we present the basic logic for perception, and define the concept of two properties being "undistinguished" (and the related notions of their being "indistinguishable" and "incompatible"). We illustrate the notion through two examples – having to do with limited visual acuity and perspective, respectively – and show that our definition avoids a traditional paradox. In Section 3 we then augment the language to incorporate belief, and discuss how the basic model can be specialized to capture various interactions between perception and belief; at the end of the section we address a more sophisticated version of the paradox of indistinguishability.

Throughout this article we treat only the single-agent case, but the extension to multiple agents is straightforward.

## 2 Perception and indistinguishability

In this section we introduce the concepts of perception and perceptual indistinguishability. We will introduce an operator ***P*** to capture perception, but we must be careful in what we aim to capture with it. We intend that a formula $P\psi$ is to be read as "$\psi$ follows from **the agent's** perceptions". Though we could have interpreted it simply as "the agent perceives $\psi$", this would force us to distinguish between what is perceived and the result of "interpreting" the perception (e.g perceiving a retinal image of a car vs. perceiving a car, which in a **sense** is an interpretation of the retinal image). This distinction is a problematic one (cf. [14]), and technically inconvenient as well; our choice allows us to treat ***P as*** a modality akin to belief and to make sense of iterated applications of the operator (for one or multiple agents.) Nonetheless, just as the distinction between explicit and implicit belief has proven important [10, 4], a notion of "immediate perceptual information" would be worthy of investigation; we do not pursue it in this article.

Perception does not in general provide complete information about what is perceived, in the sense that it may fail to distinguish between different states of the world. This notion of perceptual indistinguishability has often led, as already mentioned, to paradoxes. We need a relation that is reflexive and symmetric, but which is not transitive. The intuition we will try to capture in defining it is that two propositions are perceptually undistinguished when the perception of one does not make the other incompatible with the agent's perception; and conversely, if they are distinguished, then the perception of one rules out the other.

### 2.1 The basic formal model

The basic logic of perception is a standard modal logic. Given a set P of primitive propositions, the syntax of the logic is defined recursively as usual: any primitive proposition is a formula in the language, the special logical symbol *false* is in it, and if $\psi$ and $\phi$ are formulas, then so are $\neg\psi$ and $\psi$ A $\phi$. The usual abbreviations are used: $\psi \vee \phi = \neg(\neg\psi \wedge \neg\phi)$; $\psi \supset \phi = \neg(\psi$ A $\neg\phi)$; $\psi \equiv \phi = \psi \supset 4$ A $\phi \supset \psi$; and $true = \neg false$.

We adopt the standard possible worlds semantics, and treat ***P*** as a modal operator for perception. Given our interpretation of ***P,*** it is natural to require that it satisfies at least the axioms of KD45 [1]. Formally, a *Kripke structure for perception* is a tuple $M = (W, I, \pi)$, where $W$ *is* a set of worlds; $I$ is a serial, transitive and Euclidean relation on $W$, the accessibility relation for perception; and $\pi$ is a truth assignment, i.e. a function mapping each primitive proposition into the set of worlds in which the proposition is true, with the restriction $\pi(false) = \emptyset$. The semantics for the language is defined as follows.

$M, s \models p$ *if* $s \in \pi(p)$, for any primitive proposition $p \in$ P.

$M, s \models \neg\psi$ if $M, s \not\models \psi$.

$M, s \models \psi \wedge \phi$ if $M, s \models \psi$ and $M, s \models \phi$.

$M, s \models P\psi$ if $M, t \models \psi$ for every $t$ such that (s, $t$) $\in$ I.

The following is the standard (sound and complete) axiomatization of KD45:

A 1. All tautologies of propositional calculus.

A2. $(P\psi \wedge P(\psi \supset \phi)) \supset P\phi$.

A3. $\neg P(false)$.

A4. $P\psi \supset PP\psi$.

A5. $\neg P\psi \supset P\neg P\psi$.

Rl. From $\psi$ and $\psi \supset \phi$ infer $\phi$.

R2. From $\psi$ infer $P\psi$.

Thus, for example, if $\psi$ follows from your perceptions then it follows from your perceptions that $\psi$ follows from your perceptions. For readability, we will simply say that if the agent perceives $\psi$ then she perceives that she perceives $\psi$, etc. Some trivial results due to this treatment of perception are: the agent cannot perceive both a proposition and its negation (i.e. $\neg(P\psi \wedge P\neg\psi)$ is valid in any Kripke structure for perception); the agent perceives the conjunction of two propositions exactly when she perceives both propositions $(P\psi \wedge P\phi \equiv P(\phi \wedge \psi)$ is valid); and if the agent perceives $\psi$ then she perceives $\psi \vee \phi$ for any formula $\phi$ $(P\psi \supset P(\psi \vee \phi)$ is valid).

## 2.2 Indistinguishability

The intuition is that the agent cannot distinguish perceptually between two formulas when perceiving one is incompatible with perceiving the negation of the other. This intuition can be captured quite directly in our framework. However, rather than directly define the property of being perceptually ***indistinguishable,*** we ***first*** define the concept of two formulas being perceptually ***undistinguished at a world*** (this finer grain definition will turn out to be important). For this purpose we introduce a new operator $UD$ (for undistinguished).

***Definition*** **1** $UD(\psi, \phi) =_{def} \neg(P\psi \wedge P\neg\phi) \wedge \neg(P\phi \wedge P\neg\psi)$

Thus, if $\psi$ is undistinguished from $\phi$ and $\psi$ is perceived, then $\phi$ is compatible with the agent's perceptions; and ***vice versa.*** Conversely, they are distinguished if perceiving one entails that the negation of the other follows from the agent's perceptions. We can also say that two formulas are ***indistinguishable*** if they are undistinguished in every world in the structure and that they are ***perceptually incompatible*** if the perception of one entails that the negation of the other follows from the agent's perceptions. In the following definitions, let $M = (W, I, \pi)$ be a Kripke structure for perception, $s \in W$, and $\psi$ and $\phi$ two formulas.

Definition **2** ***$\psi$ and $\phi$ are*** undistinguished ***in*** $M$, $s$ ***iff***

$$M, s \models UD(\psi, \phi)$$

***and*** distinguished ***otherwise.***

Definition **3** $\psi$ ***and*** $\phi$ are indistinguishable ***in M iff***

$$M, s \models UD\ (\psi, \phi)\ \textit{for every}\ s \in W$$

***and*** distinguishable ***otherwise.***

Definition **4** $\psi$ ***and*** $\phi$ ***are*** perceptually incompatible ***in A4 iff***

$$M, s \models (P\psi \supset P\neg\phi) \wedge (P\phi \supset P\neg\psi)\ \textit{for every}\ s \in W$$

***and*** compatible ***otherwise.***

In the philosophical literature, it has been suggested that the concept of distinguishability leads to paradoxes, and several clever formulations have been devised to avoid those. For example, drawing on Fine's [5], in [13] Parikh proposes ***a*** logic of ***vague predicates.*** The paradox is essentially the heap paradox, and hinges on the transitivity of intistinguishability. To use an example due to Parikh, if the human eye can not discern a change in color as a result adding one drop of black paint into a gallon of white paint, and if indistinguishability is transitive, then adding two drops, three drops or 100 gallons of black paint would not result in discernible difference either, which is clearly nonsensical.'

It is not hard to see that with this definition the paradox never arises in our system: The operator $UD$ is reflexive and symmetric, but not transitive:

Proposition 1 ***(reflexivity)*** $UD\ (\psi, \psi)$ ***is valid.***

Proposition 2 ***(symmetry)*** $UD\ (\psi, \phi) \equiv UD\ (\phi, \psi)$ ***is valid.***

Proposition **3** ***(nontransitivity)*** $UD\ (\psi, \phi)$ A $UD\ (\phi, \varphi)$ A $\neg$ $UD\ (\psi, \varphi)$ ***is satisfiable.***

(As a counterexample to transitivity, consider a world in which $\neg P\phi$ and $\neg P\neg\phi$ are true (i.e there is no perceptual information about $\phi$), and in which $P\psi$ is true. Then $UD\ (\psi, \phi)$ and $UD\ (\phi, \neg\psi)$ are true, but $UD\ (\psi, \neg\psi)$ is false.)

We also have the following properties:

Proposition **4** $P\psi$ A $\neg$ $UD\ (\psi, \varphi) \supset P\neg\varphi$ *is valid.*

Proposition 5 $P\phi$ A $UD\ (\phi, \varphi) \supset \neg\ P\neg\ \varphi$ ***is valid.***

Thus, the perception of a proposition immediately rules out all other propositions which are perceptually distinguished from it; whereas propositions that are undistinguished from it remain compatible with the agent's perceptions. We will see in section 3.3 how an incorrect encoding of these two properties lies at the root of a paradox associated with the notion of indistinguishability.

Propositions (1)-(5) constitute the basic properties that we may expect of any operator for perceptual indistinguishability. Some other properties of $UD$ , which suggest that the operator is well-behaved, are given at the end of this section. First, however, we illustrate the definitions with two concrete examples.

---

**There is a more elaborate version of the paradox, due to Davis, which involves belief (or knowledge) as well as perception. We will return to it at the end of section 3.**

Example 1 ***(Sensor tolerance.)*** Consider a real-valued sensor accurate to within $\pm 2$, i.e. such that any two values differing by more than 2 points in the scale can be perceptually distinguished. Let's represent the values the sensor may take in some convenient form, e.g. as $T = x$ for some real number x; and let $val(T = x) = x$ be a metalinguistic function which gives us the value of the measure denoted by each proposition "$T = x$". This situation can be captured by using two types of axiom, one for stating the situations in which values are undistinguished $(i)$, and one for stating when they are distinguished (ii).

Axioms of type (i) are instances of the schema

$$UD(\psi, \phi)$$

for each $\psi, \phi$ such that $val(\psi) - 2 \leq val(\phi) \leq val(\psi) + 2$. It is easy to check that this is consistent with the reflexivity and symmetry of $UD$. Given these axioms, we can use proposition (5) to derive, e.g., $\neg P \neg (T = 8)$ from $P(T = 10)$ and $UD(T = 10, T = 8)$, thus obtaining that $T = 8$ is compatible with the perception $P(T = 10)$.

Axioms of type $(ii)$ consist of instances of the schema

$$P\psi \supset \neg UD(\psi, \phi)$$

for any $\psi, \phi$ such that $val(\phi) < val(\psi) - 2$ or $val(\psi) + 2 < val(\phi)$. This is sufficient to rule out all values incompatible with a particular reading. For example, $P(T = 10)$ implies $\neg UD(T = 10, T = 7)$; using proposition $(4)$, we obtain $P \neg (T = 7)$.

It should be noted that type (i) axioms will not always be adequate. For suppose we have an additional sensor, with identical accuracy, for the same property, and that the readings of each sensor are, respectively, $T = 10$ and $T = 11$. Then $P(T = 11)$ and $\neg UD$ (T = 11, T = 8) imply $P \neg (T = 8)$, in contradiction with our previous conclusion. In this case, one possibility is simply to remove type $(i)$ axioms; type (ii) axioms will still allow the agent to rule out all propositions distinguished from the perceived values. Another, quite interesting possibility is to interpret $P$, by analogy with the "all I know" operator [6, 11], as an "all I perceive" operator. If all I perceive is $T = 10$, then I can legitimately conclude $\neg P \neg (T = 8)$, as suggested in the example; but if I also perceive $T = 11$, the conclusion would no longer follow. Thus, the required semantics would be non-monotonic. □

Many variations of this example are possible. Our next example shows how perceptual indistinguishability can be modeled as a function of the agent's "perspective".

Example 2 ***(Parallax.)*** Suppose that the ability of the agent to perceptually distinguish between various propositions in a given domain depends on its physical position. For example, looking at a portion of a line from a slanted perspective allows for less resolution than looking at that same portion from a position which gives you a front view of it. To make things concrete, assume that the agent's resolution is a function of the angular separation between two points, as seen from her position, that the domain is a bidimensional surface, which we will represent by Cartesian coordinates, and that the goal is to determine the position of a mark on its x-axis. Specifically, suppose that there is some angular value $\alpha$ such that the agent can distinguish any two points separated by an angle greater than

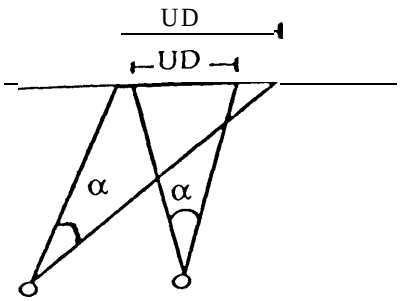


Figure 1: Example 2

$\alpha$ with origin in the agent's position (see Figure 2). The language includes two disjoint sets of propositional symbols, S and M, where symbols in S will be written $am\_at(s)$ and stand for the agent's position s in the plane, and symbols in M, written A4 = x, stand for the position of the mark in the x-coordinate. Given three points i, $j$, $k$ in the plane, $|\widehat{ijk}|$ represents (in the metalanguage) the magnitude of the angle with origin in i and whose two edges pass through j and $k$ respectively. If $\theta \in$ S has the form ***am-at(s),*** then $val(\theta)$ = s is the position in the plane denoted by 8; if $\psi \in$ M is $M$ = x then $val(\psi) = (x, 0)$ is the position of the mark in the x-axis represented by $\psi$. In a similar fashion as in the previous example, we add the axiom (of type (i)):

$$am\_at(s) \supset UD\ (\psi, \phi),$$

for any $\psi, \phi \in$ M and ***am-at(s)*** $\in$ S such that $val(am\_at(s))$ = $s$, $val(\psi) = j$, $val(\phi) = k$, and $|\widehat{sjk}| \leq \alpha$. Again, it is easy to check that this is consistent with the reflexivity and symmetry of $UD$ . Thus, two positions of the mark are undistinguished if the angle with origin in the agent's position passing through the two positions is smaller than $\alpha$; as before, we obtain that values for the position of the mark which are undistinguished from its perceived value (as seen from the agent's position), are compatible with the agent's perceptions. Similarly, we add the type (ii) axiom

$$\textbf{am-at(s)} \wedge P\psi \supset \neg\ \textbf{UD}\ (\psi, \phi)$$

for any $am\_at(s) \in$ S and $\psi, \phi \in$ M such that $val(am\_at(s))$ = $s$, $val(\psi)$ = $j$, $val(\phi) = k$, and $|\widehat{sjk}| > \alpha$, and the agent can rule out all positions of the mark which are separated from the perceived position by more than $\alpha$, taking the the agent's own position as origin. □

It should be obvious that these examples can be elaborated to include arbitrarily complex mathematical or symbolic conditions. ***One*** of these elaborations is worth noticing: We may

decide to integrate the results of various sensors as if they were only one. Thus, if the agent in our previous example had an stereoscopic vision system, the set of axioms could be modified to deal with her improved resolution directly. In general, this procedure may allow us to make stronger assertions about compatibility: if all sensors for a given domain are integrated into one, type ***(i)*** axioms can be used without worrying about the problems arising from multiple sensors for these axioms. More generally, it is in principle possible to enrich our language as much as we want so that the metalinguistic conditions that we have been using are expressible in the object language. As a simple elaboration of example 1, we could introduce a new connective $\prec$ in the language such that $\psi \prec \phi$ iff $val(\psi) < val(\phi)$. Then we could use axiom schemas such as, say,

$$\neg UD\ (\psi, \phi) \wedge \psi \prec \phi \prec \varphi \supset \neg UD\ (\psi, \varphi),$$

which together with proposition 5 gives us:

$$P\phi \wedge \neg UD\ (\psi, \phi) \wedge (\psi \preceq \phi \preceq \mu) \supset P\neg\mu,$$

and similarly for other axioms dealing with the ordering of propositions. There is no reason in principle why such enrichments could not go as far as incorporating, say, a full-fledged interval arithmetic, though obviously we would have to leave the propositional realm.

Some other properties of ***UD*** are listed next. A formula is perceptually undistinguished from its negation exactly when there is no perceptual information about it, it is distinguished from truth iff its negation is perceived and distinguished from false iff it is perceived. We also have that $\psi$ is undistinguished from both *4* and *-4* iff either there is no perceptual information about $\psi$ or there is no perceptual information about *4;* that $\psi$ cannot be simultaneously distinguished from both a formula and its negation; and that two formulas are undistinguished exactly in case their negations are undistinguished. Finally, if $(\psi \vee 4)$ is not perceived, then $\psi$ and *4* are undistinguished (or, equivalently, if $\psi \wedge \phi$ is compatible with what is perceived, then they are undistinguished.) Each of these properties is intuitively plausible, which seems to indicate that ***UD*** is a well-behaved operator.

Proposition 6 ***The following sentences are all valid:***
$UD\ (\psi, \neg\psi) \equiv \neg P\psi \wedge \neg P\neg\psi$
$\neg$ ***UD*** $(\psi,$ ***true)*** $\equiv P\neg\psi$
$\neg UD\ (\psi,$ ***false)*** $\equiv P\psi$
$UD\ (\psi, \phi) \wedge UD\ (\psi, \neg\phi) \equiv [\neg(P\psi \vee P\neg\psi) \vee \neg(P\phi \vee P\neg\phi)]$
$UD\ (\psi, \phi) \vee UD\ (A\text{-}4)$
$UD\ (\psi, \phi) \equiv UD\ (\neg\psi, \neg\phi)$
$\neg P(\psi \vee \phi) \supset UD\ (\psi, \phi)$

### *2.3* Accurate and observant agents

As mentioned in the introduction, various agent types can be defined to capture varying assumptions about the domain of application. A number of such types are discussed in section 3. Most definitions refer to the relationship between perception and belief, a topic

we have not yet addressed. However, two agent types are defined in terms of the relationship between perception and objective facts in the world. Since these definitions are independent of the belief modality, they are given here.

We first consider ***observant*** agents, those from whose perception nothing escapes. Formally, the agent is said to be observant if the following schema is valid:

$$\psi \supset P\psi,$$

Similarly, the agent is said to ***be accurate*** if the schema

$$P\psi \supset \psi$$

is valid.

Note that the axiom schema for the accurate agent is simply the schema T for the operator ***P.*** The schema corresponding to the observant agent is clearly very strong, unlikely to be applicable to agents in a complex environment, since nothing in it escapes perceptually from the observant agent. Yet it is possible to think of simple agents for which it makes sense, such as e.g. a device continuously and accurately monitoring some variable and whose model of the world consists exclusively in the possible values the variable can take. It is also possible to define agent types in terms of the validity of the respective schemas with respect to specific subsets of propositions, in which case we could think of certain "sensory modalities" being modelled by the axiom for observant agents.

The first fact to notice is the following:

Proposition 7 ***An observant agent is accurate.***

For the observant agent, furthermore, ***UD*** is transitive, and thus becomes an equivalence relation.

Proposition ***8 For the observant agent, the following sentences are valid:***
$UD(\psi, \mathbf{4}) \wedge UD(\phi, \varphi) \supset UD(\psi, \varphi)$.
$\boldsymbol{UD}(\psi, \phi) \equiv (P\psi \equiv \boldsymbol{P4})$.
$UD(\psi, \mathbf{4}) \equiv (\psi \equiv \phi)$.

For the accurate agent, the state of the world cannot be distinguished from its perceived state.

Proposition 9 ***For the accurate agent,*** $P\psi \wedge \neg UD(\psi, \mathbf{4}) \supset \mathbf{-4}$ ***is valid.***

What these properties suggest is that the most interesting uses of ***UD*** arise, as could be expected, for agents which are neither observant nor accurate.

## 3 Incorporating belief

We are now interested in what the agent should believe, given what she perceives. It might be tempting to make the assumption that anything perceived is also believed. However, in (common) cases where the sensors might not be completely accurate, and in which the

agent might know this, the assumption would be inappropriate. Sometimes the converse assumption is more adequate: one doesn't believe in anything without having perceived it. Keeping this in mind, we begin with the most general case, in which the relation between perception and belief is unconstrained, and which can be captured in a straightforward manner. We will then identify a number of special cases with further properties.

Formally, we add an operator $\boldsymbol{B}$ to our vocabulary to represent belief, and extend the language so that $B\psi$ is a formula whenever $\psi$ is a formula. A ***Kripke structure for perception and belief*** is a tuple $M = (W, \boldsymbol{R}, \boldsymbol{I}, \pi)$, where IV, $\boldsymbol{I}$ and $\pi$ are as before and $\boldsymbol{R}$ is a serial, transitive and Euclidean relation over $W$. The semantics of the $B$ operator are given as usual by:,

$$M, s \models B\psi \text{ if } M, t \models \psi \text{ for every } t \text{ such that (s, } t) \in \boldsymbol{R}.$$

The system BP0 consists of all the axioms and rules of PO together with:

A4. $(B\psi \wedge B(\psi \supset \phi)) \supset B\phi$.

A5. $\neg B(\boldsymbol{false})$.

A6. $B\psi \supset BB\psi$.

A7. $\neg B\psi \supset B\neg B\psi$.

R3. From $\psi$ infer $B\psi$.

The following result should be obvious:

Proposition 10 ***BP0 is a sound and complete axiomatization with respect to the class of Kripke structures for perception and belief.***

### 3.1 Agent types

By placing additional restrictions on I or jointly on $\boldsymbol{I}$ and $\boldsymbol{R}$ it is possible to capture different ways in which perceptions can be related to the agent's beliefs and to the environment. There are a number of basic dimensions to consider. In the previous section we appealed to the relationship between perception and the environment in defining the accurate and observant agents. In addition we may now consider the introspective abilities of the agent, the accuracy of the agent's beliefs about her perceptions and the confidence that the agent places in her perceptions. The first table defines the various agent types along these three dimensions in terms of the validity of certain (not necessarily independent) axiom schemas, and the second table identifies certain classes of Kripke structures for perception in terms of restrictions on $I$ and $\boldsymbol{R}$. The correspondence between the axiom schemas and the corresponding classes of structures is given in the soundness and completeness results in the next theorem.

Theorem 1 ***BP0 + 1 (respectively + 2, + 3, + 4, + 5, + 6, + 7, + $7_2$, + $7_3$, + 8, + $8_2$, + 9, + $9_2$) is a sound and complete axiomatization with respect to the class of Kripke structures for perception and belief*** $M_{observant}$ (respectively $M_{accurate}$, $M_{positive-intro}$, $M_{negative-intro}$, $M_{weak\text{-}positive\text{-}intro}$, $M_{weak\text{-}negative\text{-}intro}$, $M_{confident}$, $M_{confident2}$, $M_{confident3}$, $M_{cautious}$, $M_{cautious2}$, $M_{skeptical}$, $M_{skeptical2}$.)

| *Agent Type* | *Axiom schema* |
|---|---|
| 0 bservan t agent | 1. $\psi \supset P\psi$ |
| Accurate agent | 2. $P\psi \supset \psi$ |
| Positively introspective | 3. $P\psi \supset BP\psi$ |
| Negatively introspective | 4. $\neg P\psi \supset B\neg P\psi$ |
| Weakly positively introspective | 5. $B\neg P\psi \supset \neg P\psi$ |
| Weakly negatively introspective | 6. $BP\psi \supset P\psi$ |
| Perceptually confident | 7. $BP\psi \supset B\psi$ |
| Perceptually confident2 | $7_2$. $P\psi \supset B\psi$ |
| Perceptually confident$_3$ | $7_3$. $B(P\psi \supset \psi)$ |
| Cautiously confident | 8. $BP\psi \supset \neg B\neg\psi$ |
| C a u t i o u s l y confident$_2$ | $8_2$. $P\psi \supset \neg B\neg\psi$ |
| Skeptical ("show-me") | 9. $\neg BP\psi \supset \neg B\psi$ |
| Skeptical2 | $9_2$. $\neg P\psi \supset \neg B\psi$ |

| *Class of structures* | *Restriction* |
|---|---|
| $M_{observant}$ | 1. **Vs,** $t, (s, t) \in I$ iff s = *t* |
| $M_{accurate}$ | **2.** *vs,* $(s, s) \in I$ |
| $M_{positive\text{-}intro}$ | 3. $\forall s$, *t,* u, if (s, *t)* $\in R$ and *(t,* u) $\in I$ then $(s, u) \in I$ |
| $M_{negative\text{-}intro}$ | **4.** Vs, *t,* u, if $(s, t) \in R$ and (s, u) $\in I$ then (t, $u$) $\in I$ |
| $M_{weak-positive-intro}$ | 5. $\forall s \exists t$ such that $(s, t) \in R$ and **Vu,** if $(t,$ u) $\in I$ *then* $(s, u) \in I$ |
| $M_{weak\text{-}negative\text{-}intro}$ | 6. $\forall s$, u, if (s, u) $\in I$ then $\exists t$ such that (s, *t)* $\in R$ *and* $(t, u) \in I$ |
| $M_{confident}$ | i'. $\forall s, u$, if $(s, u) \in R$ then $\exists t$ such that (s, *t)* $\in R$ and $(t, u) \in I$ |
| $M_{confident2}$ | $7_2$. **Vs,** *t,* if (s, *t)* $\in R$ then (s, *t)* $\in I$ |
| $M_{confident3}$ | $7_3$. $\forall s$, *t,* if (s, *t)* $\in R$ then *(t, t)* $\in I$ |
| $M_{cautious}$ | 8. $\forall s \exists t, u$ such that (s, *t)* $\in R$ and $(t, u) \in I$ and $(s, u) \in R$ |
| $M_{cautious2}$ | $8_2$. $\forall s \exists t$ such that (s, *t)* $\in R$ and (s, *t)* $\in I$ |
| $M_{skeptical}$ | 9. **Vs,** u, if $\exists t$ such that (s, *t)* $\in R$ and $(t,$ u) $\in I$, then $(s,$ u) $\in R$ |
| $M_{skeptical2}$ | $9_2$. Vs, *t,* if $(s, t) \in I$ then (s, *t)* $\in R$ |

The introspective abilities of the agent with respect to perception refer to the agent's beliefs about her perceptions (as opposed to the propositions perceived), and give rise to two basic agent types, corresponding to positive and negative introspection of perceptions. If an agent has ***positive introspection of perception,*** then whenever the ***agent*** perceives $\psi$, she believes that she perceives it. There are situations in which this is not the case, since an agent may not be aware that she perceives something; a good deal of perceptual information is often ignored or "filtered out". For example, the ambient noise in a room may go unnoticed though we would hesitate to say that it is not perceived. In this sense, positive introspection can be interpreted as the agent's awareness of her perceptions. ***Negative introspection of perception,*** in contrast, rules out believing that you have had a perception that in fact you have not had. For agents with both types of introspection, the respective schemas can be strengthened into equivalences.

Proposition 11 ***For an agent with positive and negative introspection*** *of* ***perception,*** $(P\psi \equiv BP\psi)$ ***and*** $(\neg P\psi \equiv B\neg P\psi)$ ***are valid*** *schenzas.*

The next two types of agent refer to the accuracy of the agent's beliefs about her perceptions. The ***weakly positively introspective agent*** believes that she has not had a perception only if she in fact has not had it; and the ***weakly negatively introspective agent*** believes that she has had a perception only if she has in fact had it. Agents with both types of weak introspection have only correct beliefs about what perceptions they have had. The choice of names is justified because each of them is strictly weaker than its analogously named introspective agent. Obviously, each half of proposition 11 holds for an appropriate combination of weak and standard introspection, though neither holds in general for an agent with both types of weak introspection.

Agents may also differ in the level of confidence they place on their perceptions. Again we distinguish two basic types, a ***confident agent*** who believes in her perceptions and an agent who, more cautiously, simply refuses to believe in the falsity of what she perceives ***(cautiously confident agent).*** In both cases, there are several reasonable ways (distinguished by subscripts) of characterizing the agent. For the confident agent (with no subscript), the agent believes in the truth of the propositions that she believes she has perceived (including by the way the case in which she believes this falsely, as it may happen in the absence of weakly negative introspection). Thus, in this version, which is the preferred one, it is not sufficient to perceive $\psi$ in order for the confident agent to believe that $\psi$. It ***is*** sufficient for the confident2 agent, but its corresponding schema can be obtained from schema 5 by adding positive introspection, thus allowing for a clearer separation between perception and belief. Finally, the last version of the confident agent is given by axiom schema $5_3$, which is simply a stronger version of 5. It can perhaps be read as identifying confidence with the belief that one is accurate. Note also that the two types of cautiously confident agents are related in exactly the same way as the two first versions of the confident agent, namely: the second version can be obtained from positive introspection and the preferred version of the cautious agent.

Finally, we consider ***skeptical agents,*** who refuse to believe in something unless they (believe that) they have perceived it. The two versions are related again by positive introspection. More generally, we can capture some of the dependencies between schemas for different types of agents in proposition 12. The list is not intended to be exhaustive, and in fact it is easy to see some other dependences which are obvious consequences from those listed.

Proposition 12

***An agent with positive introspection*** *of* ***perception has weakly positive introspection*** *of per-* ***ception.***

***An agent with negative introspection*** *of* ***perception has weakly negative introspection*** *of per-* ***ception.***

***An observant and confident2 agent has positive introspection*** *of* ***perception.***

***An*** $observant$ ***and*** $confident_2$ ***agent has negative introspection*** *of* ***perception.***

***A confident3 agent is a confident agent.***

***A confident agent is cautiously confident.***

***A confident2 agent is cautiously*** $confident_2$.
***A confident agent with positive introspection of perception is*** $confident_2$.
***A cautiously confident agent with positive introspection of perception is cautiously confident;!.***
***A confident2 agent with weakly negative introspection of perception is confident.***
***A cautiously confident2 agent with weakly negative introspection of perception is cautiously confident.***
***A skeptical, agent with positive introspection of perception is a skeptical agent.***
***A skeptical agent with weakly negative introspection of perception*** is $skeptical_2$.

### 3.2 Belief and indistinguishability

In section 2.3 we noted that in the case of both accurate and observant agents, the $UD$ operator has strong additional properties. This is true also of the new agent types, though in a less extreme fashion, corresponding to the less extreme nature of the assumptions made about such types.

We leave it to the reader to explore the effects on $UD$ of the various forms of confidence. We only the note the interaction between introspection and distinguishability:

Proposition **13** ***For an agent with positive introspection of perception,***

$$\neg UD\,(\psi, \phi) \supset B \neg UD\,(\psi, \phi)$$

***is valid.***

Proposition **14** ***For an agent with negative introspection of perception,***

$$UD\,(\psi, \phi) \supset B\,UD\,(\psi, \phi)$$

***is valid.***

These results can be read as follows. Negative introspection can be seen as awareness of the lack of perceptual information, and positive introspection as awareness of certain positive perceptual information. The indistinguishability of two propositions can also be seen as a lack of information, and the negatively introspective agent is aware of it; similarly, the positively introspective agent is aware of the positive perceptual information given by two propositions being distinguished. Combining both results, an agent with both positive and negative introspection of perception has only correct beliefs (and has all correct beliefs) about his ability to perceptually distinguish between formulas.

Proposition 15 ***For the agent with positive and negative introspection of perceptions,***

$$UD\,(\psi, \phi) \equiv B\,UD\,(\psi, \phi)$$
$$\neg UD\,(\psi, \phi) \equiv B \neg UD\,(\psi, \phi).$$

***are valid.***

Interestingly, none of the results in this section hold for weakly introspective agents. Furthermore, though the weakly negatively introspective agent is the converse of the positively introspective agent, the converse of proposition 13 does not hold for the former; and similarly, the converse of proposition 14 does not hold for the weakly positively introspective agent. Thus, whereas the agent with weakly negative introspection cannot have incorrect beliefs about the perceptions she has had, she may have incorrect beliefs about the propositions that she has perceptually distinguished. The reason is that though the agent may not incorrectly believe that she has had a perception, she may fail to believe (be unaware) that she has had some particular perception. This loss of information is reflected in her failure to have only correct beliefs about what she can distinguish. With more information, the problem dissapears, as the following proposition makes clear.

Proposition 16 ***For the agent with weakly negative introspection of perception,***

$$BP\psi \wedge B\neg UD(\psi,\phi) \supset \neg UD(\psi,\phi)$$

***is valid.***

### 3.3 The paradox of indistinguishability, revisited

In section 2.2 we mentioned the paradox of indistinguishability, and how it does not arise in our framework due to the intransitivity of the $UD$ operator. However, in [3] E. Davis articulated a more sophisticated version of the paradox, which involves both perception and belief (or, in his case, knowledge). In his words:

> This property of indistinguishability can lead to the following paradox: Let a, b and c be three states of the world such that a is indistinguishable from b and b from c, but a is not indistinguishable from c. Suppose the real state of the world is a. Then the agent can see that the world is not in state c. If he knows enough about his own perceptions, then, even though he cannot see that the state of the world is not b, he can infer it, since, if the world were in b then he could not see that the world is not in state c.

Davis reconstructs this argument within a formal system, and then proceeds to proposes two ingenious, though somewhat complex, modifications which escape the paradox.

Although intuitively compelling, this argument (like all those leading to paradoxes) contains a certain slight of hand, which recasting in our formal system will expose. The reason Davis was able to present a formal version of the argument in the first place is that he implicitly made quite strong assumptions: that perception entails belief; that perception, and hence belief, are accurate (indeed, he talks of knowledge rather than belief); that agents have strong introspective capabilities; and that indistinguishability is defined relative to the whole structure and not at a world. His two solutions amount to relaxing these implicit assumptions in different ways.

In our general framework, which does not assume a certain agent type, the argument does not go through. It can be derived, but only by making strong assumptions along the way. Here is a derivation faithful to the English text, annotated by the assumptions made:

1. ***B(a*** ∨ ***b*** **v c)** (given)
2. a (given)
3. $UD\ (a, b)$ (given)
4. $UD\ (b,$ ***c) (given)***
5. $\neg\ UD$ ***(a, c)*** (given)
6. $\forall x,\ y.\ x$ A $\neg UD\ (x,$ ***y)*** ⊃ P-y (assumption)
7. $B(\forall x,$ y- $x \wedge \neg UD$ *(x, y)* ⊃ *P-y)* (assumption)
8. $\forall x, y.$ x A $UD\ (x,$ *y)* ⊃ *-P-y* (assumption)
9. $B(\forall x,\ y.$ x A ***UD*** (x, y) ⊃ $\neg P \neg y)$ (assumption)
10. $P \neg c$ (From 2+5+6)
11. ***BP-c*** (From 10, assuming ***Pp*** ⊃ $BPp$)
12. ***B-b*** (From 9+11, assuming $UD\ (p, q) \supset BUD\ (p, q)$)
13. ***B-c*** (From 10, assuming ***Pp*** ⊃ ***Bp)***
14. ***Ba*** (From 1+12+13)

Some of the assumptions made along the way correspond to certain agent types[2] and it can be argued that it is desirable to escape the paradox even for those types. However, assumptions 6 and 8 (and their counterparts, 7 and 9) are completely unfounded. They should be rather replaced by propositions 4 and 5, instantiated here, respectively, to:

$$Pa \text{ A } \neg \text{ } UD \text{ } (a, \text{ } c) \supset P\text{-}c. \tag{1}$$

$$Pb \text{ A } UD \text{ } (b, \text{ } c) \supset \text{-}P\text{-}c. \tag{2}$$

The rationale for these two propositions is that even if ***c*** is ***perceptually*** distinguished from a, c cannot be ruled out perceptually unless the agent ***perceives a.*** If the world is in state a but the agent perceives ***b,*** then ***Pb*** together with (2) implies *-P-c,* contrary to assumption 6; and if we had *b* and ***Pa,*** then (1) implies *P-c,* contrary to 8. For example, if the temperature is 8 and we have a thermometer accurate to within 2 (i.e values are distinguished when they differ by more than this margin), it is quite possible to have a reading of 10, which does not allow us to rule out a temperature of 12. It is only if the ***reading*** (as opposed to the actual temperature) differs from a value by more than the margin of error that we can rule out that value. In general, therefore, the paradox does not arise in our framework.

There is at least one case which assumptions 6 and 8 are valid, and that is the case of the observant agent. Indeed, we have the following properties:

[2]Specifically, **the assumptions made in lines 11 and 13 correspond to positive introspection of perception and confidence$_2$, respectively; the assumption *in* line 12 is a *consequence* of negative introspection of perception.**

Proposition 17 ***For the observant agent,*** $\psi$ A $\neg UD\ (\psi, \phi) \supset$ ***P-4*** is valid.

Proposition 18 ***For the observant agent,*** $\psi$ A ***UD*** $(\psi, \phi) \supset \neg P \neg \phi$ ***is valid.***

However, this hardly poses a paradox, since (as shown in section 2.3) for observant agents ***UD*** is transitive, rendering the given facts 3, 4 and 5 mutually inconsistent.

# 4 Concluding remarks

We have presented a formalism to reason about perception and belief, a heretofore underexplored area of investigation. We have explored possible relations between an agent's perceptions and beliefs, and between them and the environment, under different assumptions, likely to be applicable in different setups. No type of agent among those defined is likely to be usable in all scenarios, and we have mentioned some of the circumstances in which different agent types should be used.

In addition, we have defined the notion of perceptual indistinguishability in our framework in a way which we think is simple and intuitive. The operator can be easily used to formalize a wide class of examples in which perception is involved, and allows us to capture a limited notion of "perceptual perspective," without incurring in paradoxes.

The formalism we have presented seems natural and well-behaved, but more research is needed on applying, whether in general formulation of commonsense reasoning or in more specific applications. It would also be interesting to study extensions of the language which incorporate more specific domain structure (such as a metric space), and relate them to perceptual indistinguishability. Temporal and multi-agent aspects are also of interest, as well as the problem of updating the beliefs of the agent on the light of perceptually acquired information.

# Appendix

Proof of theorem 1. We prove the soundness part first.

(Observant: $\psi \supset P\psi$.) Obvious.

(Accurate: $P\psi \supset \psi$.) If $M$, s $\models P\psi$ then for every $t$ such that (s, $t$) $\in I$,$M$, $t \models \psi$. By restriction 2, (s, s) $\in I$, so $M$, s $\models \psi$.

(Positive introspection of perception: $P\psi \supset BP\psi$.) Assume M, s $\models P\psi$, i.e. M, $t \models \psi$ for every $t$ such that (s, $t$) $\in$ I. We need to show: $M$, u $\models \psi$ for every $t$, u such that (s, $t$) $\in \boldsymbol{R}$ and $(t, u) \in I$. But $(s, t) \in \boldsymbol{R}$ and $(t$, u) $\in$ I imply $(s, u) \in$ I, by restriction 3, and therefore $M, u \models \psi$.

(Negative introspection of perception: $\neg P\psi \supset B\neg P\psi$.) If M, s $\models \neg P\psi$ then there exists u such that (s, u) $\in$ I and$M$, u $\models \neg\psi$. Let ul be one such $u$. We need to show: for every $t$ such that $(s, t) \in \boldsymbol{R}$ there exists v such that $(t, v) \in$ I and $M$, v $\models \neg\psi$. But if (s, $t$) $\in \boldsymbol{R}$ and (s, $u1$) $\in I$ then (t, $u1$) $\in$ I, by restriction 4, so $u1$ is such v.

(Weakly positive introspection of perception: $B\neg P\psi \supset \neg P\psi$.) Suppose $M, s \models B\neg P\psi$. Then for every $t$ such that (s, $t$) $\in \boldsymbol{R}$, there exists u such that $(t$, u) $\in$ I and $M$, u $\models \neg\psi$.

By restriction 5, there exists $t$ such that $(s,t) \in R$ and such that for all u if $(t,$ u$) \in$ I then $(s,$ u$) \in I$. It follows that for some u such that (s, $u$) $\in r$, $M, u \models \neg\psi$.

(Weakly negative introspection of perception: $BP\psi \supset P\psi$.) Suppose $M$, s $\models BP\psi$ and (s, $u$) $\in I$. Restriction 6 implies that there exists $t$ such that (s, $t$) $\in R$ and $(t, u) \in I$. Thus, $M, t \models P\psi$, so $M, u \models \psi$.

(Confident: $BP\psi \supset B\psi$.) Completely analogous to the previous case.

(Confident$_2$: $P\psi \supset B\psi$.) If $M$, s $\models P\psi$ then for every $t$, (s, $t$) $\in I$ implies $M, t \models \psi$. But since (s, $t$) $\in R$ implies $(s,t) \in$ I, by restriction $7_2$, we have that (s, $t$) $\in R$ implies $M, t \models \psi$.

(Confident$_3$: $B(P\psi \supset \psi)$.) If $(t,t) \in I$, then we have $M, t \models P\psi \supset \psi$. By restriction $7_3$, for ***every*** s and $t$ such that (s, $t$) $\in R$ we have that$(t, t) \in$ I. It follows that for every such s and $t$, $M, t \models P\psi \supset \psi$, and therefore $M$, s $\models B(P\psi \supset \psi)$.

(Cautiously confident: $BP\psi \supset \neg B\neg\psi$.) If M, s $\models BP\psi$ then for every $t$ and u such that (s, $t$) $\in R$ and (t, u) $\in$ I we have $M$, u $\models \psi$. By restriction 8, there exists v satisfying these two conditions such that in addition (s, v) $\in R$. Since $M$, v $\models \psi$, $M$, s $\models \neg B\neg\psi$.

(Cautiously confident$_2$: $P\psi \supset \neg B\neg\psi$.) If $M$, s $\models P\psi$ then for every $t$, (s, $t$) $\in I$ implies $M, t \models \psi$. By restriction $8_2$, there exists $t$ such that$(s, t) \in I$ and (s, $t$) $\in R$, so there exists $t$ such that (s, $t$) $\in R$ and $M, t \models \psi$.

(Skeptical: $\neg BP\psi \supset \neg B\psi$.) If $M, s \models \neg BP\psi$ then there exists $t$, u such that (s, $t$) $\in R$ and $(t,$ u$) \in$ I and such that $M$, u $\models \neg\psi$. By restriction 9, $(s,$ u$) \in R$, so $M$, s $\models \neg B\psi$.

(Skeptical$_2$: $\neg P\psi \supset \neg B\psi$.) The proof is identical to that for the confident;! agent, reversing the roles of $R$ and $I$.

The completeness results that follow do not assume $BP0$ but only the subset of it formed by propositional logic and the distributivity axioms (i.e. Al, A2 and A4), in addition to the applicable rules of inference. The system formed from these axioms and all rules together with a particular axiom a will be named BP$a$ (so for example BP1 stands for this system augmented with axiom 1), and $\vdash_{BPa}$ will be used to denote the consequence relation in each BP$a$.

To prove the completeness of an axiomatic system $\Sigma$ with respect to a given class of structures $\mathcal{M}$, we need to show that if a sentence is valid in $M$ (i.e. valid in every $M \in M$) then it is provable in $\Sigma$, or, equivalently, that if it is C-consistent then it is satisfiable in some $M \in \mathcal{M}$. The techniques we will use are fairly standard, so to avoid repetition we will go only over the details of the proof that are specific to our results (cf. [1, 7] for a more detailed discussion). By Lindembaum's lemma, every C-consistent set of sentences (where $\Sigma$ is any axiomatic system including Al and Rl, i.e. propositional logic) can be extended to a maximal C-consistent set, one such that the addition of a new formula would make it inconsistent ([1]). F***or*** any given class of structures ***M,*** we construct a canonical structure $M^c \in \boldsymbol{M}$ such that there is a world $s_v$ corresponding to every maximal C-consistent set $V$. Then we show that $M^c, s_v \models \varphi$ iff $\varphi \in V$. This suffices to prove the completeness of $\Sigma$, as it is easy to see. By Liendembaum's lemma, if $\varphi$ is C-consistent then it is contained in some maximal C-consistent set $V$ and therefore $M^c, s_v \models \varphi$; but this **implies that $\varphi$ is** satisfiable in $M$, as we want to show. To construct the canonical structure $M^c$ for the class of Kripke structures for perception, we proceed as follows. Given a set $V$ of formulas, let

$V/B = \{\varphi \mid B\varphi \in V\}$ and V/P = $\{\varphi \mid P\varphi \in V\}$. Then $M^c$ = (S, *R*, *I*, $\pi$), where

$$S = \{s_v \mid V \text{ is a maximal consistent set}\}$$
$$s_v \in \text{n-(p) iff } p \in V$$
$$R = \{(s_v, s_w) \mid V/B \subseteq W\}$$
$$I = \{(s_v, s_w) \mid V/P \subseteq W\}.$$

Now we can show that $M^c, s_v \models \varphi$ if and only if $\varphi \in V$ by induction on the structure of $\varphi$; in addition, we can show that if all substitution instances of the schemas in BP0 are valid in $M^c$ then the relations *R* and I of $M^c$ must both be serial, transitive and euclidcan. It follows that $M^c$ is in the class of Kripke structures for perception, which proves the completeness of BP0 with respect to this class of structures. We prove the other cases in more detail next.

(Observant: BP1.) We have seen that for the observant agent, $\varphi \equiv P\varphi$ (since the observant agent is also accurate). Therefore $V/P = \{\varphi \mid P\varphi \in V\} = \{\varphi \mid \varphi \in V\} = V$. It follows that if (s,, $s_w$) $\in I$, then $V \subseteq W$. But $V \not\subset W$, since otherwise *W* would be inconsistent by the maximality of *V*. Therefore if (s,, $s_w$) $\in I$ then $V = W$, *so* $s_v = s_w$.

(Accurate: BP2.) By axiom 2, if $P\varphi \in V$, then $\varphi \in$ V-Therefore, if $\varphi \in V/P$ then $\varphi \in V$, so $V/P \subseteq V$ and $(s_v, s_v) \in I$.

(Positive introspection of perception: BP3.) Assuming the validity of axiom 3, we have to show that if $(s_v, s_w) \in R$ and (s,, s,) $\in I$, then $(s_v, s_x) \in$ I, or equivalently that if $V/B \subseteq W$ and $W/P \subseteq$ X then $V/P \subseteq$ X. By axiom 3, if $P\varphi \in V$, then $BP\varphi \in V$, and therefore $P\varphi \in W$ and $\varphi \in X$. But then $V/P \subseteq X$.

(Negative introspection of perception: BP4.) Assuming the validity of axiom 4 and that $(s_v, s_w) \in R$ ($V/B \subseteq W$) and $(s_v$, s,) $\in$ I ($V/P \subseteq$ X), we have to show that $(s_w, s_x) \in I$, i.e. that $W/P \subseteq$ X. Suppose, for indirect proof, that $W/P \not\subseteq$ X. Then there exists a formula $\varphi$ such that $P\varphi \in W$ and $\neg\varphi \in$ X; since $V/P \subseteq$ X, $M^c, s_v \not\models P\varphi$, and by the maximality of $V$, $M^c, s_v \models \neg P\varphi$. By axiom 2, $M^c, s_v \models B\neg P\varphi$, so $M^c, s_w \models \neg P\varphi$ and $P\varphi \notin W$, in contradiction with the hypothesis. Therefore $W/P \subseteq X$.

(Weakly positive introspection of perception: BP5.) For *V*, *W* and X BP5-consistent sets, we need to show that for every $V$ there exists a $W$ such that $W/B \subseteq V$ and such that for every X, if $W/P \subseteq$ X then $V/P \subseteq$ X. We show first that $V/B \cup \{P\psi \in V\}$ *is* BPS-consistent. Suppose not. Then there exist formulas $B\sigma_1, \ldots, B\sigma_n$ and $P\phi_1, \ldots, P\phi_m$ such that $B\sigma_i \in V$ ($1 \leq i \leq n$) and $P\phi_j \in$ X ($1 \leq$ j $\leq$ **nz),** and such that

$$\vdash_{BP5} \sigma_1 A \ldots A \sigma_n A P\phi_1 A \ldots A P\phi_m \supset \textit{false},$$

or equivalently, such that

$$\vdash_{BP5} \sigma_1 \wedge \ldots \wedge \sigma_n \supset \neg(P(\phi_1 \wedge \ldots \wedge \phi_m).$$

Using R3, we get

$$\vdash_{BP5} B[\sigma_1 \text{ A.. . } \wedge \sigma_n \supset \neg P(\phi_1 \wedge \ldots \wedge \phi_m)],$$

which implies

$$\vdash_{BP5} B\sigma_1 \text{ A.. . A } B\sigma_n \supset B\neg P(\phi_1 \text{ A.. . A } \phi_m).$$

Since each $B\sigma_i \in V$, $B\neg P(\phi_1 \wedge \ldots \wedge \phi_m) \in V$, and by axiom 5, $\neg P(\phi_1 \wedge \ldots \wedge \phi_m) \in V$. And since each each $P\phi_j \in V$, $P(\phi_1 \wedge \ldots \wedge \phi_m) \in V$, a contradiction. Therefore $V/B \cup \{P\psi \in V\}$ is BP5-consistent, and by Liendembaum's lemma there exists a maximal BP5-consistent set $W$ such that $(V/B \cup \{P\psi \in V\}) \subseteq W$. This means in particular that $V/B \subseteq W$ and that $\{P\psi \in V\} \subseteq \{P\psi \in W\}$, so $V/P \subseteq W/P$, and thus for every X, if $W/P \subseteq X$ then $V/P \subseteq X$, as we wanted to show.

Some of the following results assume the following lemma. For a maximal C-consistent set $V$, define $V/P^* = \{\neg P\neg\varphi \mid \varphi \in V\}$.

**Lemma 1** *If $V$ and $W$ are two maximal C-consistent sets, then $V/P \subseteq W$ iff $W/P^* \subseteq V$.*

***Proof.*** From left to right, suppose $V/P \in W$, so that if $P\varphi \in V$ then $\varphi \in W$, and suppose $\varphi \in W$. Then $\neg\varphi \notin W$, so $P\neg\varphi \notin V$, and by the maximality of $V$, $\neg P\neg\varphi \in V$. Thus, if $\varphi \in W$ then $\neg P\neg\varphi \in V$, so $W/P^* \subseteq V$. From right to left, suppose $W/P^* \subseteq V$, so that if $\varphi \in W$ then $\neg P\neg\varphi \in V$, and suppose that $P\varphi \in V$. Then $\neg P\varphi \notin V$, so $\neg\varphi \notin W$, and by the maximality of $W$, $\varphi \in W$. Thus, if $P\varphi \in V$ then $\varphi \in W$, so $V/P \subseteq W$. □

(Weakly negative introspection of perception: BPG.) Assume (s,, s,) ∈ I, so that $V/P \subseteq$ X. We have to show that there exists $s_w \in$ S such that (s,, $s_w$) ∈ $R$ and (s,, $s_x$) ∈ $I$, or equivalently, that there exists a maximal BP&consistent set $W$ such that $V/B \subseteq W$ and $W/P \subseteq$ X. By lemma 1, it suffices to show that there exists $W$ such that $V/B \cup X/P^* \subseteq W$. By Liendenbaum's lemma, such $W$ exists if $V/B$ u $X/P^*$ is BP&consistent. Suppose on the contrary that this union is inconsistent. Then there exist formulas $\sigma_1, \ldots, \sigma_n$ and $\neg P\neg\phi_1, \ldots, \neg P\neg\phi_m$ such that $B\sigma_i \in V$ $(1 \leq i \leq n)$ and $\phi_j \in$ X $(1 \leq j \leq m)$, and such that

$$\vdash_{BP6} \sigma_1 \wedge \ldots \wedge \sigma_n \wedge \neg P\neg\phi_1 \wedge \ldots \wedge \neg P\neg\phi_m \supset \mathit{false},$$

or equivalently, such that

$$\vdash_{BP6} \sigma_1 \wedge \ldots \wedge \sigma_n \supset (P\neg\phi_1 \vee \ldots \vee P\neg\phi_m).$$

But this last formula implies

$$\vdash_{BP6} \sigma_1 \wedge \ldots \wedge \sigma_n \supset P(\neg\phi_1 \vee \ldots \vee \neg\phi_m),$$

which in turn is equivalent to

$$\vdash_{BP6} \sigma_1 \wedge \ldots \wedge \sigma_n \supset P\neg(\phi_1 \wedge \ldots \wedge \phi_m).$$

Using R3, we get

$$\vdash_{BP6} B[\sigma_1 \wedge \ldots \wedge \sigma_n \supset P\neg(\phi_1 \wedge \ldots \wedge \phi_m)],$$

which implies

$$\vdash_{BP6} B\sigma_1 \wedge \ldots \wedge B\sigma_n \supset BP\neg(\phi_1 \wedge \ldots \wedge \phi_m).$$

Since each $B\sigma_i \in V$, $BP\neg(\phi_1 \wedge \ldots \wedge \phi_m) \in V$, and by axiom 6, $P\neg(\phi_1 \wedge \ldots \wedge \phi_m) \in V$. By the hypothesis that $V/P \subseteq$ X, it follows that $\neg(\phi_1 \wedge \ldots \wedge \phi_m) \in X$. But since each

$\phi_j \in X, (\phi_1 \wedge \ldots \wedge \phi_m) \in X$, a contradiction. Therefore $V/B \cup X/P^*$ is BPG-consistent, as we wanted to show.

(Confident: BP7.) The proof is analogous to the previous case. Assuming $V/B \subseteq X$, we have to show that $V/B \cup X/P^*$ is consistent. If it is not, then there exist formulas $\sigma_1, \ldots, \sigma_n$ and $\neg P\neg\phi_1, \ldots, \neg P\neg\phi_m$ such that $B\sigma_i \in V$ $(1 \leq i \leq n)$ and $\phi_j \in X$ $(1 \leq j \leq m)$ and such that

$$\vdash_{BP7} B\sigma_1 \wedge \ldots \wedge B\sigma_n \supset BP\neg(\phi_1 \wedge \ldots \wedge \phi_m).$$

Since each $B\sigma_i \in V$, $BP\neg(\phi_1 \wedge \ldots \wedge \phi_m) \in V$, and by axiom 7, $B\neg(\phi_1 \wedge \ldots \wedge \phi_m) \in V$. By the hypothesis that $V/B \subseteq X$, it follows that $\neg(\phi_1 \wedge \ldots \wedge \phi_m) \in X$. But since each $\phi_j \in X$, $(\phi_1 \wedge \ldots \wedge \phi_m) \in X$, a contradiction.

(Confident$_2$: BP7$_2$.) If $(s_v, s_w) \in R$, then by axiom $7_2$ if $P\varphi \in V$ then $B\varphi \in V$ and therefore $\varphi \in W$; so $V/P \subseteq W$ and therefore $(s_v, s_w) \in I$

(Confident$_3$: BP7$_3$.) If $(s_v, s_w) \in R$, then by axiom 73, $B(P\varphi \supset \varphi) \in V$, so $(P\varphi \supset \varphi) \in W$. Thus, if $P\varphi \in W$ then $\varphi \in W$, so $W/P \subseteq W$ and $(s_w, s_w) \in I$.

For the cautiously confident agent, we need the following lemma. For a maximal $\Sigma$-consistent set $V$, define $V/BP = \{\varphi \mid BP\varphi \in V\}$.

Lemma *2 If V and X are two maximal C-consistent sets, then $V/BP \subseteq X$ iff there exists some maximal C-consistent set W such that $(s_v, s_w) \in R$ and $(s_w, s_x) \in I$.*

***Proof.*** From left to right, assume $V/BP \subseteq X$. We need to show that there exists a maximal C-consistent set $W$ such that $V/B \subseteq W$ and $W/P \subseteq X$, or by lemma 1 such that $V/B \cup X/P^* \subseteq W$. By Liendembaum's lemma, it suffices to show that $V/B \cup X/P^*$ is consistent. Suppose not. Then there exist formulas $\sigma_1, \ldots, \sigma_n$ and $\neg P\neg\phi_1, \ldots, \neg P\neg\phi_m$ such that $B\sigma_i \in V$ $(1 \leq i \leq n)$ and $\phi_j \in X$ $(1 \leq j \leq m)$, and such that

$$\vdash_{\Sigma} \sigma_1 \wedge \ldots \wedge \sigma_n \wedge \neg P\neg\phi_1 \wedge \ldots \wedge \neg P\neg\phi_m \supset \textit{false}.$$

We can then derive as before:

$$\vdash_{\Sigma} B\sigma_1 \wedge \ldots \wedge B\sigma_n \supset BP\neg(\phi_1 \wedge \ldots \wedge \phi_m).$$

Since each $B\sigma_i \in V$, $BP\neg(\phi_1 \wedge \ldots \wedge \phi_m) \in V$, and by the hypothesis that $V/BP \subseteq X$, it follows that $\neg(\phi_1 \wedge \ldots \wedge \phi_m) \in X$. But since each $\phi_j \in X$, $(\phi_1 \wedge \ldots \wedge \phi_m) \in X$, a contradiction.

From right to left, assume there exists $W$ such that $(s_v, s_w) \in R$ and $(s_w, s_x) \in I$. Then $V/B \subseteq W$ and $W/P \subseteq X$. Then if $BP\varphi \in V$, then $P\varphi \in W$, so $\varphi \in X$. Thus, $V/BP \subseteq X$.□

(Cautiously confident: BP8.) We have to show that that for every $s_v \in S$ there exist $s_w, s_x \in S$ such that $(s_v, s_w) \in R$ and $(s_w, s_x) \in I$ and $(s_v, s_x) \in R$. By lemma 2, we know that there exists $W$ such that $(s_v, s_w) \in R$ and $(s_w, s_x) \in I$ if and only if $V/BP \subseteq X$. Thus, we only need to show that there exists a maximal BP8-consistent set X such that $V/BP \subseteq X$ and $V/B \subseteq X$, for which in turn jt suffices to show that $V/BP \cup V/B$ is BP8-consistent. Suppose not. Then there exist formulas $\sigma_1, \ldots, \sigma_n$ and $\phi_1, \ldots, \phi_m$ such that $B\sigma_i \in V$ $(1 \leq i \leq n)$ and $BP\phi_j \in V$ $(1 \leq j \leq m)$, and such that

$$\vdash_{BP8} \sigma_1 \wedge \ldots \wedge \sigma_n \wedge \phi_1 \wedge \ldots \wedge \phi_m \supset \textit{false},$$

or equivalently, such that

$$\vdash_{BP8} \sigma_1 \wedge \ldots \wedge \sigma_n \supset \neg(\phi_1 \wedge \ldots \wedge \phi_m).$$

By rule R3, it follows that

$$\vdash_{BP8} B\sigma_1 \text{ A } \ldots \text{ A } B\sigma_n \supset B\neg(\phi_1 \text{ A} \ldots \text{ A } \phi_m),$$

and since each $B\sigma_i \in V$, $B\neg(\phi_1 \text{ A} \ldots \wedge \phi_m) \in V$, and using axiom 8, $\neg BP(\phi_1 \text{ A} \ldots \wedge \phi_m) \in \boldsymbol{V}$. But since each $BP\phi_j \in V$ and $BP\phi_1 \text{ A} \ldots \text{ A } BP\phi_m$ implies $BP(\phi_1 \boldsymbol{A} \ldots \boldsymbol{A} \phi_m)$, it follows that $BP(\phi_1 \wedge \ldots \wedge \phi_m) \in \boldsymbol{V}$, which is impossible. Therefore $\boldsymbol{V/BP} \cup \boldsymbol{V/B}$ is BP&consistent.

(Cautiously confident$_2$: BP8$_2$.) We need to show that there exists a maximal BP8$_2$-consistent set $W$ such that $\boldsymbol{V/B} \subseteq \boldsymbol{W}$ and V/P $\subseteq W$. Again, it suffices to show that $\boldsymbol{V/B} \cup \boldsymbol{V/P}$ is BP&-consistent. Suppose not. Then there exist formulas $\sigma_1, \ldots, \sigma_n$ and $\phi_1, \ldots, \phi_m$ such that $P\sigma_i \in \boldsymbol{V}$ $(1 \leq i \leq n)$ and $B\phi_j \in \boldsymbol{V}$ $(1 \leq j \leq m)$, and such that

$$\vdash_{BP8_2} \sigma_1 \text{ A } \ldots \text{ A } \sigma_n \text{ A } \phi_1 \text{ A } \ldots \text{ A } \phi_m \supset \text{false},$$

or equivalently, such that

$$\vdash_{BP8_2} \sigma_1 \wedge \ldots \text{ A } \sigma_n \supset \neg(\phi_1 \text{ A} \ldots \text{ A } \phi_m).$$

By rule R2,

$$\vdash_{BP8_2} P\sigma_1 \text{ A} \ldots \text{ A } P\sigma_n \supset P\neg(\phi_1 \text{ A} \ldots \text{ A } \phi_m).$$

Since each $P\sigma_i \in \boldsymbol{V}$, $P\neg(\phi_1 \text{ A} \ldots \text{ A } \phi_m) \in V$, and using axiom 8$_2$, $\neg B(\phi_1 \text{ A} \ldots \text{ A } \phi_m) \in V$. But since each $B\phi_j \in \boldsymbol{V}$, we have $B(\phi_1 \text{ A} \ldots \vee \phi_m) \in \boldsymbol{V}$, a contradiction.

(Skeptical: BP9.) We need to show that for every $\boldsymbol{V}$ and X, if there exists $W$ such that $\boldsymbol{V/B} \subseteq \boldsymbol{W}$ and $\boldsymbol{W/P} \subseteq \boldsymbol{X}$, then $\boldsymbol{V/B} \subseteq X$. **By** lemma 2, it suffices to show that if $\boldsymbol{V/BP} \subseteq$ X then $\boldsymbol{V/B} \subseteq$ X. So assume that $\boldsymbol{V/BP} \subseteq X$. Then *if* $B\psi \in V$, by axiom 9, $BP\psi \in \boldsymbol{V}$, so $\psi \in$ X. It follows that $\boldsymbol{V/B} \subseteq X$.

(Skeptical$_2$: BP9$_2$.) Analogous to the proof for the confident;! agent.

Proof of selected propositions.

Proposition 7. ***An observant agent is accurate.***

***Proof.*** $\neg\psi$ implies $P\neg\psi$ for an observant agent, which in turn implies $\neg P\psi$. By contraposition, $P\psi \supset \psi$.□

Proposition ***8 For the observant agent, the following are valid sentences:*** $UD(\psi, \phi)$ A $UD(\phi, \varphi) \supset UD(\psi, \varphi)$

$UD(\psi, \phi) \equiv (P\psi \equiv P\phi)$

$UD(\psi, \phi) \equiv (\psi \equiv \phi)$

***Proof.*** We prove transitivity only. The other two are an inmediate consequence of the fact that, for the observant agent, we have $\psi \equiv P\psi$ and $P\psi \equiv \neg P\neg\psi$.

$UD(\psi, \phi)$ $\boldsymbol{A}$ $UD(\phi, \varphi) \equiv (P\psi \supset \neg P\neg\phi)$ A$(P\phi \supset \neg P\neg\psi)$ A$(P\phi \supset \neg P\neg\varphi)$ A$(P\varphi \supset \neg P\neg\phi)$. Suppose $P\psi$. Then $\neg P\neg\phi$, which for the observant agent implies $\phi$ and therefore $P\phi$, which in turn implies $\neg P\neg\varphi$. So $P\psi \supset \neg P\neg\varphi$. Suppose now $P\varphi$. Then $\neg P\neg\phi$, which

as before implies $P\phi$, which implies $\neg P\neg\psi$, so $P\varphi \supset \neg P\neg\psi$. But $(P\psi \supset \neg P\neg\varphi) \wedge (P\varphi \supset \neg P\neg\psi) \equiv$ ***UD*** $(\psi, \varphi)$.□

Proposition 11. ***For an agent with positive and negative introspection of perception,*** $(P\psi \equiv BP\psi)$ ***and*** $(\neg P\psi \equiv B\neg P\psi)$ ***are valid schemas.***

***Proof.*** For the first one, from left to right, this is simply positive introspection. From right to left, $BP\psi$ implies $\neg B\neg P\psi$, which by negative introspection implies $P\psi$. For the second one, from left to right, this is negative introspection. From right to left, $B\neg P\psi$ ***implies*** $\neg B\neg\neg P\psi$, i.e. $\neg BP\psi$, which by positive introspection implies $BP\psi$.□

Proposition **12.** We prove only the less obvious dependencies.

***An observant and confident2 agent has positive introspection of perception.***

***Proof.*** $P\psi \supset PP\psi$ for an observant agent, which implies ***B*** $P\psi$ for $\text{confident}_2$. □

***An observant and confident2 agent has positive introspection of perception.***

***Proof.*** $\neg P\psi \supset P\neg P\psi$ for observant, which implies $B\neg P\psi$ for $\text{confident}_2$. □

***A*** $\textit{confident}_2$ ***agent with weakly negative introspection of perception is confident.***

***Proof.*** $BP\psi \supset P\psi$ by weakly negative introspection, and $P\psi \supset B\psi$ for $\text{confident}_2$.

***A cautiously confident;! agent with weakly negative introspection of perception is cautiously confident.***

***Proof.*** $BP\psi \supset P\psi$ by weakly negative introspection, which implies $\neg B\neg\psi$ for cautiously confident;!. □

## References


[1] B. F. Chellas. ***Modal Logic: An Introduction.*** Cambridge University Press, 1980.

[2] P. R. Cohen and H. Levesque. Intention is choice with commitment. ***Artificial*** *Intelligence*, 42(3), 1990.

[3] E. Davis. Solutions to a paradox of perception with limited acuity. In ***Proceedings of the First International Conference on Knowledge Representation and*** *Reasoning*, ***1989.***

[4] R. Fagin and J. Y. Halpern. Belief, awareness and limited reasoning. ***Artificial Intelligence,*** 34( 1):39–76, 1988.

[5] K. Fine. Vagueness, truth and logic. *Synthèse*, 30:301–324, ***1975.***

[6] J.Y. Halpern and Y. Moses. Towards a theory of knowledge and ignorance. In ***Proceedings of*** *AAAI* ***Workshop on Nonmonotonic Logic,*** pages 125-143, 1984.

[7] J.Y. Halpern and Y. Moses. A guide to the modal logics of knowledge and belief: Preliminary draft. In ***Proceedings of the 9th International Joint Conference on Artificial Intelligence, 1985.***

[8] J . Hintikka. ***Knowledge and Belief.*** Cornell University Press, 1981.

[9] K. Konolige. ***A Deduction Model of Belief*** Pittman Research Notes in Artificial Intelligence, 1986.

[10] H.J. Levesque. A logic of implicit and explicit belief. In ***Proceedings of AAAI,*** pages 198–202, Austin, TX, 1984.

[11] H.J. Levesque. All I know: A study in autoepistemic logic. ***Artificial Intelligence,*** 42:263–309, 1990.

[12] A.K. Mackworth and R. Reiter. A logical framework for depiction and image interpretation. ***Artificial Intelligence,*** 4 1:125–155, 1990.

[13] R. Parikh. The problem of vague predicates. In R.S. Cohen and M.W. Wartofsky, editors, ***Language, Logic and Method.*** Reidel Publishing Company, 1983.

[14] W. 0. Quine. ***Word and Object.*** MIT Press, 1964.

[ 15] Yoav Shoham. Agent-oriented programming. Technical Report TR STAN-CS-1335 (revised), Stanford University, 1990.

[16] M.Y. Vardi. On epistemic logic and logical omniscience. In ***Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge, 1986.***