Seeing is Believing

John Bell and Zhisheng Huang
Applied Logic Group
Computer Science Department
Queen Mary and Westfield College
University of London, London E1 4NS, UK
{jb, huang}@dcs.qmw.ac.uk

Abstract

In this paper we present a formal common sense theory of perception and belief. We begin with a logical analysis of perception, and present formal semantics for a modal perception operator. We then consider the relationship between perception and belief, and in particular the question of when perception should lead to belief change. Our theory is intended to apply to both to high-level, mental, perception in humans and to perception at many levels in artificial agents, but especially at the level of the symbolic interface between a vision system and a belief system.

1 Introduction

It is a commonplace that seeing is believing. When asked to justify a belief, people often reply "I saw it with my own eyes". Indeed, people are often not prepared to believe something unless they have seen it for themselves. Perceptions also occupy a place of special importance in Epistemology, being considered as fundamental, incorrigible, etc. However, there is always the possibility that perceptions are illusory. Consequently, philosophers, notably Berkeley, have questioned whether the common sense world that we know of as a result of our perceptions really exists; see e.g. [1] for a discussion.

It is not the task of AI perception systems to entertain sceptical doubts of a philosophical nature. Their task is rather to produce the high-level perceptions which the philosophers take as their starting point. Systems of this kind are thus mainly concerned with the task of interpreting sensor data. For example a vision system might start with a two dimensional array of pixel data supplied by its camera, and its task is to interpret the image, inferring edges, then surfaces, then objects, etc., and perhaps finally producing a high-level symbolic description; see e.g. [7, 11] for an introduction.

However, once this task is accomplished, there remains the question of the interaction between the agent's vision system and its belief system, and it is here that the problem of illusion arises. Should the agent believe what it sees? While philosophers may entertain sceptical doubts because of possiblity of illusion, common sense demands that, other things being equal, we should believe what we see, and AI systems should do likewise. So if an agent's perceptions are consistent with its beliefs, the agent should believe what it sees. Thus, following Pollock [9], Musto and Konolige [8] suggest that seeing is normally believing. Consequently, in their outline of a formal theory of perception, they give an appropriate rule formulated in autoepistemic logic. But what if the agent's perceptions and beliefs conflict? Should the agent regard its perception as illusory and ignore it, or should it regard its beliefs as being mistaken and revise them in order to accommodate its perception? Rather than argue dogmatically for one of these alternatives, which may be appropriate in some circumstances but not in others, we suggest a flexible, pragmatic, solution : it should depend on the comparative reliability of the perceptions and the beliefs in question. If the perception is more reliable, the agent should revise its beliefs to accommodate it. But if the beliefs are more reliable, the perception should be considered illusory and the agent's beliefs should remain unchanged.

In this paper we develop a formal common sense theory of perception and its interaction with belief along the lines suggested. Our theory is expressed in the language $\mathcal{CA}$ [2], a many-sorted first-order modal language with explicit reference to time points. This has been extended to include the preference operator of ALX [4, 5] and Lewis's counterfactual conditional [6]. The relevant features of $\mathcal{CA}$ will be introduced as they are needed. As is common in philosophy, we often discuss sight (or vision) but the remarks are applicable to the remaining senses (modes of perception). The rest of this paper is organised as follows. In the next section we propose a logical analysis of perception. We then outline our theory of belief and belief revision in Section 3. The interactions between perception and belief are then formalised in the concluding section.

2 Perception

We begin with the question of what *is* perceived, or what are the "objects of perception". We suggest, with Musto and Konolige [8] that we perceive *perceptual attributes*, or simply *perceptibles*. They define these as follows:

> A perceptual attribute is any property of physical objects whose presence or absence can be detected directly by sensing. Examples of perceptual attributes are *spatial* attributes like shape, size, and location, and *non-spatial* attributes, like color, temperature, and weight. Attributes like "belonging-to" or "flammable", whose presence or absence cannot be judged simply by sensing, are not included in the set of perceptual attributes.

We suggest that the formal counterparts of perceptual attributes are atomic sentences of the form $r(u_1, \ldots, u_n)$ where, r is a relation symbol and the u_i are terms. Thus we suggest that we do not perceive perceptual attributes which correspond to sentences which are logically more complex. One cannot for example *perceive* that there is not a block on the table : one's perception cannot be of the form $\neg r(u_1, \ldots, u_n)$. Perceiving that there is not a block on the table is equivalent to perceiving that there is not a blue whale, or a pink elephant, or an invisible white rabbit, on the table. As we will see in the sequel, the fact that we don't perceive a block on the table leads us to believe that there is not one. It might be thought that the perception that two objects, o and o', are distinct involves both negation and identity, however, once again, perceiving that o is not o' is equivalent to perceiving that o is not a pink elephant, etc. So we suggest that perceptions of this kind are best represented by means of a binary $Distinct$ predicate. Similarly, one does not perceive conjunctions, disjunctions, implications or quantifications. For example, one may see a red block on the table and one may see a yellow block on the table, but one does not see the conjunction of these perceptions. Rather one's perception leads one to believe in the conjunction. Similarly, one does not perceive that all of the blocks on the table are red. Rather one perceives individual red blocks, comes to believe that there are no other blocks, and then infers that all the blocks are red. Thus, while perceptions are logically simple, they lead to inferences and beliefs which are logically complex. Note that we disagree with Musto and Konolige on this point. They suggest that perceptible attributes can correspond to existential statements, such as "there is a doorway", and attributive statements such as "the door is open". We suggest rather that the perceptions are "raw", pre-logical, descriptions, such as"d is an object", "d is a door", "d is open", and that the quantifiers arise as a result of inferences on the beliefs corresponding to these. As suggested above, we also disagree with their axiom: $P\alpha \wedge P\alpha' \supset P(\alpha \wedge \alpha')$, where α and α' are perceptibles and $P\alpha$ states that α is a current perception of the agent, as we claim that it is not possible to perceive logical conjunctions.

However, while perceptions don't have a logic, in the sense of the usual connectives and quantifiers, they should be *coherent*. Perceptible attributes can be classified into types, such as shape and colour, and it is impossible to perceive an object which has two different attributes of the same type at the same time. For example an object cannot be perceived as being both red and green all over at the same time. Although, of course, objects can have different attributes of the same type at different times. For example, the colour of silk can change according to the light and the angle at which you look at it.

We are now in a position to develop our formal theory. The atomic sentences of $\mathcal{CA}$ all have a temporal index. Thus $\mathcal{CA}$ contains sentences such as $OnTable(B)(3)$ which states that block B is on the table at time 3. For simplicity we will assume that time is discrete and linear. The models of $\mathcal{CA}$ are fairly complex possible-worlds structures. Each model comes equipped with an interpretation function $\mathcal{V}$ which assigns an n-ary relation to each n-ary relation symbol at each time point and each possible world. Thus, for model M, world w in M and variable assignment g:[1]

$$M, w, g \models r(u_1, \ldots u_n)(t) \textit{ iff } (u_1, \ldots, u_n) \in \mathcal{V}(r, t, w)$$

Let R be the set of all relation symbols of $\mathcal{CA}$, and let $R_P \subseteq R$ be the set of perceptual attributes. Then $P_T \in \mathcal{P}(R_P)$ is the set of *perceptual attribute types*. For example, P_T might be the set $\{\{red, green, \ldots,\}, \{square, circle, \ldots\}\}$ of perceptual attribute types, making $\{red, green, \ldots\}$ a perceptual attribute type. In order to represent perceptions we introduce the modal operator Per, intuitively, $Per(a, \alpha)(t)$ states that agent a perceives α at time t. The semantic component corresponding to the perception operator in a $\mathcal{CA}$ model M is a function $\mathcal{R}_{(Per,a,t,w,r)}$ which assigns an n-ary relation to each (n-ary) perception attribute $r \in R_P$ at each time point t at each possible world w for each agent a. Each $\mathcal{R}_{(Per,a,t,w,r)}$ is required to satisfy the following coherence condition:

> (PCT) If $(u_1, \ldots, u_n) \in \mathcal{R}_{(Per,a,t,w,r)}$ and $(u_1, \ldots, u_n) \in \mathcal{R}_{(Per,a,t,w,r')}$ and $r \neq r'$, then r and r' belong to different types in P_T, namely, for any $C \in P_T$ such that $r \in C$, $r' \notin C$.

The truth condition for the perception operator is then as follows:

$$M, w, g \models Per(a, r(u_1, \ldots, u_n))(t) \text{ iff } (u_1, \ldots, u_n) \in \mathcal{R}_{(Per,a,t,w,r)}$$

Clearly, it follows from (PCT) that an agent's perceptions are coherent.

This analysis is easily extended to allow for more than one mode of perception; the five senses in humans, and vision, sensors, etc., in robots. The perception operator would have an additional argument indicating the mode of perception. Thus $Per(a, \alpha, s)(t)$, would state that a perceives α with sense s at time t, etc.

Perceptions are related by their degree of reliability. For example, if a witness saw the accused at close hand and in broad daylight, their testimony is more reliable that if they had seen the accused on a dark night. Similarly a robot footballer's attempts to distinguish between ball and the goal might be more reliable in some lighting conditions than in others [Rix 1997]. Similarly with multiple senses, one sense may be more reliable than another; e.g. someone who has lost their spectacles might rely more on touch that sight, and a robot's sensors might be more reliable than its vision system. To reflect the differing reliability of perceptions, we introduce preferences between perceptions, resulting in *perception hierarchies*. The sentence $Pref(a, Per(a, \alpha)(t), Per(a, \alpha'))(t))(t)$ states that, at time t, agent a's perception of α is more reliable than a's perception of α'. In complex sentences the same agent and time point terms are

[1] For the sake of simplicity of presentation we will let the distinction between terms and their denotations in M given g take care of itself.

often repeated. In order to abbreviate complex sentences we will adopt the convention that a missing agent term is the same as the closest agent term to its left, and that a missing time point term is the same as the closest time point term to its right. Thus, for example, $Pref(a, Per(a,\alpha)(t), Per(a,\alpha')(t))(t)$ can be abbreviated to $Pref(a, Per(\alpha), Per(\alpha'))(t)$.

The semantics of the preference operator begin with von Wright's conjunction expansion principle [14]. According to this principle, to say that you prefer an apple to an orange is to say that you prefer situations in which you have an apple and no orange to those in which you have an orange and no apple. In possible-worlds terms this principle might be stated as follows: agent a prefers ϕ to ψ if a prefers $\phi \wedge \neg\psi$-worlds to $\psi \wedge \neg\phi$-worlds. However, this semantics of preference is too simple as it leads to the "paradoxes of conjunction and disjunction"; if ϕ is preferred to ψ then $\phi \vee \chi$ is preferred to ψ, and ϕ is preferred to $\psi \wedge \chi$. Examples of the undesirability of these properties are: "if a prefers coffee to tea then a prefers coffee or poison to tea" and "if a prefers coffee to tea then a prefers coffee to tea and a million dollars". Clearly we need to capture the ceteris paribus nature of preferences: we should compare $\phi \wedge \neg\psi$-worlds and $\psi \wedge \neg\phi$-worlds which otherwise differ as little as possible from the actual world. In order to do so we introduce the selection function from the Stalnaker-Lewis analysis of conditionals [6, 13]. Thus the function cw is of type $W \times \mathcal{P}(W) \to \mathcal{P}(W)$, and, intuitively, $cw(w, [\![\phi]\!]^M_g)$ is the set of closest worlds to w in which ϕ is true.[2] Formally, cw is required to satisfy the conditions imposed by Lewis in [6]. The agent's preferences over time are represented by the function $\succ: A \times T \to \mathcal{P}(\mathcal{P}(W) \times \mathcal{P}(W))$, which assigns a comparison relation over sets of worlds to each agent at each time point. Intuitively, for sets of worlds X and Y, $X \succ_{(a,t)} Y$ means that agent a prefers the worlds in X to the worlds in Y at time t. Preferences are required to be irreflexive and transitive. Accordingly, let $X \sqsupset_{(a,t,w)} Y$ abbreviate $cw(w, X \cap \overline{Y}) \succ_{(a,t)} cw(w, Y \cap \overline{X})$.[3] Then each $\sqsupset_{(a,t,w)}$ is required satisfy the following properties:

(irp)	$X \not\sqsupset_{(a,t,w)} X$.
(trp)	*If* $X \sqsupset_{(a,t,w)} Y$ *and* $Y \sqsupset_{(a,t,w)} Z$ *then* $X \sqsupset_{(a,t,w)} Z$.

The truth condition for preferences is then as follows:

$$M, w, g \models Pref(a, \phi, \psi)(t) \quad \textit{iff} \quad [\![\phi]\!]^M_g \sqsupset_{(a,t,w)} [\![\psi]\!]^M_g.$$

Given these semantics, we have the following axioms:

(IRP)	$\neg Pref(a, \phi, \phi)(t)$
(TRP)	$Pref(a, \phi, \psi)(t) \wedge Pref(a, \psi, \chi)(t) \to Pref(a, \phi, \chi)(t)$
(CEP)	$Pref(a, \phi, \psi)(t) \leftrightarrow Pref(a, (\phi \wedge \neg\psi), (\neg\phi \wedge \psi))(t)$

[2] As usual, $[\![\phi]\!]^M_g$ denotes the set of worlds in M in which ϕ is satisfied by g; i.e., $[\![\phi]\!]^M_g = \{w \in W : M, w, g \models \phi\}$.

[3] Where $\overline{X} = W - X$.

(IRP) and (TRP) state the irreflexivity and transitivity of preferences respectively, while (CEP) states the conjunction expansion principle.

Preferences between perceptions are required to satisfy the following realism condition:

$$(PRB) \quad \mathit{Pref}(a, Per(\alpha), Per(\alpha'))(t) \rightarrow Per(a, \alpha)(t) \wedge Per(a, \alpha')(t).$$

3 Belief

In this section, we briefly describe our theory of belief and belief revision based on dynamic belief hierarchies. The details can be found in [3].

At any point in time an agent has a set of *candidate beliefs*, a *candidate belief set*. Intuitively, a candidate belief is a proposition which the agent has good evidence for believing when considered in isolation. Accordingly a sentence of the form $CBel(a, \phi)(t)$ states that ϕ is a candidate belief of agent a's at time point t. The semantics for candidate beliefs are the standard possible-worlds semantics for belief, but indexed by agent and time point. Typically the agent considers some of its candidate beliefs to be more reliable than, or *preferable to*, others. These preferences thus define a *preference ordering* on the agent's candidate belief set. Typically the preference ordering is partial. Two candidate beliefs may also be equally reliable, in which case the agent has no preference between them, the agent regards them *indifferently*. Thus $\mathit{Pref}(a, CBel(\phi), CBel(\psi))(t)$ states that at, at time t, agent a considers the candidate belief ϕ to be more reliable than the candidate belief ψ,while $Ind(a, CBel(\phi), CBel(\psi))(t)$ states that a considers the two candidate beliefs to be equally reliable. In order that indifference on preferences is transitive, the prefence order on candidate beliefs is additionally required to be almost-connected. Then, at time t, an agent's candidate belief is a belief if it is *coherent*. Coherence is defined recursively, a candidate belief is coherent at t if it is jointly consistent with every belief which is at least as reliable as it at t. Thus:

$$Bel(a, \phi)(t) \leftrightarrow CBel(a, \phi)(t) \wedge Coherent(\phi)(t)$$

An agent's belief hierarchy at any time point should be the maximally coherent substructure of its candidate belief hierarchy at that time point. Thus we require that:

$$\begin{aligned} \mathit{Pref}(a, Bel(\phi), Bel(\psi))(t) \quad \leftrightarrow \quad & \mathit{Pref}(a, CBel(\phi), CBel(\psi))(t) \wedge \\ & Bel(a, \phi)(t) \wedge Bel(a, \psi)(t). \end{aligned}$$

Typically the agent's belief hierarchy is *dynamic*; as time progresses the agent's beliefs and the preferences among them change. However, the belief hierarchies of rational agents tend to exhibit a certain *stability*. The agent's beliefs and the preferences among them thus *persist* by default. Indeed, the belief hierarchies of rational agents tend to be *upwardly stable*; that is, the higher the belief in the hierarchy, the more it tends to remain in and maintain its relative position in it. This reflects the principle that rational agents should keep higher-level beliefs in preference to lower-level beliefs whenever coherence permits. However, if a rational agent realises that its beliefs are incoherent, the agent should revise them in order to restore coherence. In doing so the agent should retain more preferred beliefs in favour of less preferred ones wherever coherence permits. Moreover, the agent should only make those changes which are necessary in order to restore coherence. In order to represent the dynamic aspect of an agent's beliefs, we use the modal operator *Aff*, which is analogous to the *Ab* predicate of the Situation Calculus. Let Φ be a meta-variable which ranges over the non-temporal component of atomic formulas and atomic belief formulas. Then a formula $\Phi(t)$ is affected at t if its truth value at t differs from its truth value at $t+1$:

$$M, w, g \models \mathit{Aff}(\Phi)(t) \quad \mathit{iff} \quad M, w, g \models \neg(\Phi(t) \leftrightarrow \Phi(t+1)).$$

We thus have the following *persistence rule*:

$$\Phi(t) \wedge \neg \mathit{Aff}(\Phi)(t) \rightarrow \Phi(t+1) \tag{1}$$

Intuitively we are interested in models in which this schema is used from left to right only in order to reason "forwards in time" from instances of its antecedent to instances of its consequent. Typically also we want to be able to infer the second conjunct of each instance nonmonotonically whenever it is consistent to do so. For example, if we have $Bel(a,\phi)(t)$ then we want to be able to use the rule to infer $Bel(a,\phi)(t+1)$ if $\mathit{Aff}(Bel(a,\phi))(t)$ cannot be inferred. In order to enforce this interpretation, we define a prioritised form of preferential entailment [12].

Definition 3.1 *Let $A_1, \ldots, A_n$ be a partition of the atomic sentences (modal or otherwise) of n different types according to their type.*[4] *For each A_i, model M and time point t, let $M_{A_i}/t = \{\alpha_i(t') \in A_i \mid t' \leq t, M \models \alpha_i(t')\}$. Then a model M is* chronologically less defined that a model M' on the basis of the priorities $\langle A_1, \ldots, A_n \rangle$, *written $M \prec_{\langle A_1,\ldots,A_n\rangle} M'$ iff M and M' differ at most on the interpretation of $A_1, \ldots, A_n$ and there is a time point t such that:*

- *for some i such that $1 \leq i \leq n, M_{A_i}/t \subset M'_{A_i}/t$, and*
- *for all j such that $1 \leq j \leq i, M_{A_j}/t \subseteq M'_{A_j}/t$.*

[4]For example, A_1 might be the set *Rel* of all non-modal atoms $r(u_1, \ldots, u_n)(t)$, A_2 might be the set *Bel* of all belief atoms $Bel(a,\phi)(t)$, A_3 might be the set *Pref* of all preference atoms $\mathit{Pref}(a,\phi,\psi)(t)$, etc.

Definition 3.2 *A model M is an $\langle A_1, \ldots, A_n \rangle$-preferred model of a sentence ϕ if $M \models \phi$ and there is no model M' such that $M' \models \phi$ and $M' \prec_{\langle A_1,\ldots,A_n \rangle} M$. Similarly, M is an $\langle A_1, \ldots, A_n \rangle$-preferred model of a set of sentences Θ if $M \models \Theta$ and there is no model M' such that $M' \models \Theta$ and $M' \prec_{\langle A_1,\ldots,A_n \rangle} M$.*

Definition 3.3 *A set of sentences Θ* preferentially entails *a sentence ϕ given the priorities $\langle A_1, \ldots, A_n \rangle$ (written $\Theta \approx_{\langle A_1,\ldots,A_n \rangle} \phi$) if, for any $\langle A_1, \ldots, A_n \rangle$-preferred model M of Θ, $M \models \phi$.*

In the case of belief hierarchies, we are interested in $\langle CBel, Pref, Aff \rangle$-preferred models. As a result of the definitions we have:

Proposition 3.4 *Dynamic properties of beliefs and belief hierarchies.*

1. *Beliefs persist by default.*
2. *Preferences on beliefs persist by default.*
3. *Belief hierarchies persist by default.*
4. *Belief hierarchies are upwardly stable.*

4 Perception and Belief

We are now ready to consider the question of when perception should lead to a change in belief.

We begin by supposing that the agent's perception at a time point is consistent with its beliefs at that time point. Should the agent believe what it sees? While philosophers may entertain sceptical doubts because of possiblity of illusion, see e.g. [1], common sense demands that, other things being equal, we believe what we see. Thus, following Pollock [9], Musto and Konolige suggest, that seeing is normally believing. Consequently they give an appropriate rule formulated in autoepistemic logic. In our theory the appropriate rule can be stated as follows:

$$Per(a, \alpha)(t) \wedge \neg Bel(a, \neg\alpha)(t) \rightarrow Bel(a, \alpha)(t + 1). \tag{2}$$

Thus if, at time t, agent a perceives α, and a has no reason to doubt this perception, then a should believe what it perceives.

We also need a rule which, on the basis of non-perceptions, leads by default to negative beliefs:

$$\neg Per(a, \alpha)(t) \wedge \neg Bel(a, \alpha)(t) \rightarrow Bel(a, \neg\alpha)(t + 1). \tag{3}$$

Once again, the intended interpretation is that these rules are used from left to right only and that the second conjunct is inferred nonmonotonically.

Example 4.1 *If, at time* 1*, Stan perceives two blocks on the table, and he does not perceive any other blocks on the table, then, at time* 2*, Stan should believe that there are exactly two blocks on the table. Moreover, Stan should continue to believe this as long as his perceptions remain unchanged. Let* Θ *be the theory* $\{(1),(2),(3),(4),(5)\}$*, where:*

$$Per(Stan, Block(B1))(1) \wedge Per(Stan, Block(B2))(1) \wedge$$
$$Per(Stan, OnTable(B1))(1) \wedge Per(Stan, OnTable(B2))(1) \wedge$$
$$Per(Stan, Distinct(B1, B2))(1) \tag{4}$$
$$Bel(a, Distinct(x, y))(t) \rightarrow Bel(a, x \neq y)(t) \tag{5}$$

Then in the $\langle Per, Bel, Pref, Aff\rangle$*-preferred models of* Θ *it follows by* (2) *that Stan believes at time* 2 *that* $B1$ *and* $B2$ *are distinct blocks on the table, and by* (3) *that Stan believes that these are the only blocks on the table at time* 2*. Hence* $\Theta \mid\!\approx_{\langle Per,Bel,Pref,Aff\rangle} Bel(Stan, \exists_2 x(Block(x) \wedge OnTable(x)))(2)$*. As Stan's perceptions are unaffected, it follows by* (1) *that Stan has the analogous belief at time* 3*.* □

But what if the agent's perceptions and beliefs conflict? Should the agent regard its perception as illusory and ignore them, or should it regard its beliefs as being mistaken and revise them in order to accommodate its perception? Rather than argue dogmatically for one of these alternatives, we suggest a flexible, pragmatic, solution. Thus our answer is that it all depends on the comparative reliability of the perceptions and the relevant beliefs. If the perception is more reliable, the agent should revise its beliefs to accommodate it. But if the beliefs are more reliable, the perception should be considered illusory and the agent's beliefs should remain unchanged. In order to formalise this we introduce preferences between perceptions and beliefs. Thus $Pref(a, Bel(\phi), Per(\alpha))(t)$ states that agent a considers the belief ϕ is more reliable than the perception α. Similarly, $Pref(a, Per(\alpha), Bel(\phi))(t)$ states that a considers the perception α is more reliable than the belief ϕ. Heterogeneous preferences of this kind are required to satisfy the following realism condition:

$$Pref(a, Per(\alpha), Bel(\phi))(t) \vee Pref(a, Bel(\phi), Per(\alpha))(t) \rightarrow$$
$$Bel(\phi)(t) \wedge Per(\alpha)(t) \tag{6}$$

The appropriate rules can then be formulated as follows:

$$Pref(a, Per(\alpha), Bel(\neg\alpha))(t) \rightarrow Bel(a, \alpha)(t+1) \tag{7}$$
$$Pref(a, Bel(\neg\alpha), Per(\alpha))(t) \rightarrow \neg Bel(a, \alpha)(t+1) \tag{8}$$

Example 4.2 *At time* 1*, Stan believes that a statue,* S*, is white. However, it looks red to him. He notices that there is a red light shining on the statue, which would explain why the statue looks red. Consequently, he considers that the red* *appearance of the statue is illusory, and so at time* 2 *he continues to believe that the statue is white. However, at time* 2*, Stan no longer perceives the red light. In the absence of any other explanation for the red appearance of the statue, Stan considers that his perception that the statue is red is more reliable than his belief that it is white. Consequently, he revises his beliefs and so at time* 3 *he believes that the statue is red.*

In order to formalise Stan's belief about the causal properties of the red light we adopt Lewis's account of causal dependence [6]. Informally, a sentence of the form $\phi \Rightarrow \psi$ *is true at a world* w *if in each of the closest worlds to* w *in which* ϕ *is true it is also the case that* ψ *is true. Formally:*

$$M, w, g \models \phi \Rightarrow \psi \text{ iff } cw(w, [\![\phi]\!]_g^M) \subseteq [\![\psi]\!]_g^M.$$

Then causal dependence defined as follows:

$$Causes(\phi, \psi) \leftrightarrow (\phi \Rightarrow \psi) \wedge (\neg\phi \Rightarrow \neg\psi).$$

Thus ϕ *causes* ψ *if* ϕ *is normally sufficient for* ψ, $\phi \Rightarrow \psi$, *and normally necessary for* ψ, $\neg\phi \Rightarrow \neg\psi$.

Let $\Theta = \{(2), (3), (7), (8), (9), (10), (11), (12), (13)\}$, *where:*

$$Per(Stan, Red(S))(1) \wedge Bel(Stan, White(S))(1) \wedge$$
$$Per(Stan, Redlight)(1) \wedge \neg Per(Stan, Redlight)(2) \tag{9}$$
$$Bel(Stan, Causes(Redlight, Per(Red(S)))(t) \tag{10}$$
$$Bel(a, \neg(White(x) \wedge Red(x)))(t) \tag{11}$$
$$Per(a, \alpha)(t) \wedge Bel(a, \neg\alpha)(t) \wedge$$
$$\exists\alpha'(Per(a, \alpha')(t) \wedge Bel(a, Causes(\alpha', Per(\alpha)))(t)) \rightarrow$$
$$Pref(a, Bel(\neg\alpha), Per(\alpha))(t) \tag{12}$$
$$Per(a, \alpha)(t) \wedge Bel(a, \neg\alpha)(t) \wedge$$
$$\neg\exists\alpha'(Per(a, \alpha')(t) \wedge Bel(a, Causes(\alpha', Per(\alpha)))(t)) \rightarrow$$
$$Pref(a, Per(\alpha), Bel(\neg\alpha))(t) \tag{13}$$

The existential quantifiers in (12) *and* (13) *are introduced purely for convenience, and can in practice be replaced by disjunctions.*

Then in $\langle Per, Bel, Pref, Aff\rangle$*-preferred models of* Θ *it follows from* (12) *that* $Pref(Stan, Bel(\neg Red(S), Per(Red(S)))(1)$, *and thus, by* (7), *that* $Bel(Stan, White(S))(2)$. *Moreover, it follows, from* (13), *that* $Pref(Stan, Per(Red(S), Bel(\neg Red(S)))(2)$, *so it follows from* (8) *that* $Bel(Stan, Red(S))(3)$. □

Acknowledgements

This research forms part of the Ratio Project and is supported by the United Kingdom Engineering and Physical Sciences Research Council under grant number GR/L34914.

References

[1] A.J. Ayer (1976) *The Central Questions of Philosophy*, Penguin Books, London.

[2] J. Bell. Changing Attitudes. In: M.J. Wooldridge and N.R. Jennings (Eds.). *Intelligent Agents.* Post-Proceedings of the ECAI'94 Workshop on Agent Theories, Architectures, and Languages. Springer Lecture Notes in Artificial Intelligence, No. 890. Springer, Berlin, 1995. pp. 40-55.

[3] J. Bell and Z. Huang. Dynamic Belief Hierarchies, *Proceedings of the Second Workshop on Practical Reasoning and Rationality*, Manchester, April 1997.

[4] Huang, Z., *Logics for Agents with Bounded Rationality*, ILLC Dissertation series 1994-10, University of Amsterdam, 1994.

[5] Huang, Z., Masuch, M., and Pólos, L., ALX: an action logic for agents with bounded rationality, *Artificial Intelligence* 82 (1996), pp. 75-127.

[6] Lewis, D., *Counterfactuals*, Basil Blackwell, Oxford, 1973.

[7] D. Marr. *Vision.* W. Freeman and Company, New York, 1982.

[8] Musto, D., and Konolige, K., Reasoning about perception, 1993 AAAI Spring Symposium on Reasoning about Mental States, 1993, pp. 90-95.

[9] J. Pollock (1974) *Knowledge and Justification*, Princeton University Press, Princeton, New Jersey.

[10] A. Rix, (1997) *Vinnie, a Robot Footballer*, M.Sc. Project Report, Dept. of Computer Science, Queen Mary and Westfield College.

[11] S. Russell and P. Norvig (1995) *Artificial Intelligence; A Modern Approach*, Prentice Hall, New Jersey.

[12] Shoham, Y. *Reasoning About Change*, M.I.T. Press, Cambridge, Massachusetts, 1988.

[13] Stalnaker, R., A Theory of Conditionals, *Studies in Logical Theory, American Philosophical Quarterly* 2 1968, pp. 98-122.

[14] G. von Wright (1963) *The Logic of Preference*, Edinburgh University Press, Edinburgh.