Technical Report OUCS-2006-03

My beliefs about your beliefs - a case study in theory of mind and epistemic logic

Authors:

Hans van Ditmarsch and Willem Labuschagne
Computer Science, University of Otago

Status: Journal Submission

Department of Computer Science,
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

http://www.cs.otago.ac.nz/research/techreports.html

My beliefs about your beliefs — a case study in theory of mind and epistemic logic

Hans van Ditmarsch & Willem Labuschagne[*]

August 28, 2006

Abstract

We model three examples of beliefs that agents may have about other agents' beliefs, and provide motivation for this conceptualization from the theory of mind literature. We assume a modal logical framework for modelling degrees of belief by partially ordered preference relations. In this setting, we describe that agents believe that other agents do not distinguish among their beliefs ('no preferences'), that agents believe that the beliefs of other agents are in part as their own ('my preferences'), and the special case that agents believe that the beliefs of other agents are exactly as their own ('preference refinement'). This multi-agent belief interaction is frame characterizable. We provide examples for introspective agents. We investigate which of these forms of belief interaction are preserved under three common forms of belief revision.

1 Introduction

Reasoning about preferences and reasoning about degrees of belief have independent roots and have also seen interactions combining both [18, 15, 16, 37, 33, 28, 8, 13]. We assume a multi-agent framework to reason about preferences that corresponds to reasoning about degrees of belief. Given a domain of abstract states that satisfy certain factual descriptions, an agent may *prefer* one state over another one, in the sense that the one state is considered more likely than the other state. (Our notion of preference models degrees of belief, or level of confidence. This is in the Lewis [18] sense of the term. We do not identify 'preferred' with (more) 'desirable', as in the BDI logics [23] sense: a state can very well be more desirable than another state, but still considered less likely. We apologize to those readers used to the other meaning of 'preference'.) For example, given a two-agent system where operators Anne and Bill operate a control room with a fan and a light, it may be that Anne considers it more likely that the light is on, than that the light is off. Given a domain consisting of a state s where the light is on and another state s' where the light is off, we then say

[*] University of Otago, Dunedin, New Zealand, {hans,willem}@cs.otago.ac.nz

that Anne *prefers* s over s'. The relation between preferences and degrees of belief is as follows. Anne 'normally', 'tentatively', or 'defeasibly' *believes* that the light is on because this is true in all preferred states, i.e., in s. But it is not the case that she 'strongly believes', 'is convinced', or 'knows' that the light is on, because there remains a less preferred, plausible, state where the light is off, namely s'—and she may later see reason to change her beliefs. More preference distinctions result in more degrees of belief. Thus, a preference system can be seen as a partial order on a domain of abstract states, which induces a set of degrees of belief. A main point is that preferences do no *have* to be seen as between factual descriptions of states, but may also involve *epistemic* features of those states. For example (assuming a different model), given that the light is off, Anne may consider it more likely that Bill also considers it more likely that the light if off, than not. In other words, agents may have preferences about other agents' preferences. We call this *higher-order preferences.*

Apart from such features of models wherein higher-order preferences play a role, there are also deeper structural ways in which agents' preferences relate to each other. For example, Anne may *always* consider it more likely that Bill considers it more likely that the light is off, if that is also her own preference. She may in fact strongly believe that that Bill's preferences are exactly as her own. That is a frame-characterizable preference. We model three such frame-characterizable preferences: 'another agent's preferences are (exactly) similar to my own', 'the other agent has no preferences' (or I have no reason to believe so), and 'another agent's preferences are "somewhat like" my own preferences.'

The multi-modal and multi-agent framework in which we express preferences by degrees of belief, can also be used to express how our preferences change after (higher-order, iterated) belief revision. We determine which of the characterizable preferences that we distinguish are preserved under particular forms of such belief revision.

This investigation aims to explore, on the basis of a psychologically realistic foundation, logical formalizations of multi-agent systems in which the defeasible beliefs of agents include in their scope the defeasible beliefs of other agents (or of themselves). As foundation we propose the corpus of 'theory of mind' publications. This work is informed by cognitive and communicative deficiencies in agents who are immature or suffer a mental deficit of some kind, and thus is drawn from the empirical content of such branches of psychology as developmental psychology and the study of autism. The term 'theory of mind' is used here in the technical sense of the term in psychology and should not be interpreted as indicating any desire on our part to contribute to the 'philosophy of mind' discussion. The intended direction of information flow is from psychology to logic, with the potential for a measure of reverse flow. The philosophy of mind is neither the source nor the target of the information flow in this investigation, although empirical results from psychology are of course relevant for philosophers also. In the present context we are concerned with the extent to which traditional logical treatments of belief and knowledge are able to accommodate what we have learned from psychology about the way in which different naturally occurring classes of agent impute beliefs to other agents.

In Section 2 we give an overview of relevant 'theory of mind' literature. In Section 3 we introduce a multi-agent logic for reasoning about degrees of belief. In Sections 4.1, 4.2, and 4.3 we characterize three forms of higher-order belief interaction. We illustrate the different types with examples involving introspective agents. In Section 5 we determine what belief interaction is preserved under different forms of belief revision.

2 Theory of Mind

The term 'Theory of Mind' was defined by Premack and Woodruff [22] as the ability to impute mental states to oneself and to others. Clearly, it is a crucial component of social skills, since predicting what other agents will do in a situation is facilitated by having an idea of what they believe to be the case in that situation. In particular, a fully developed Theory of Mind (ToM) will recognise that other agents may have beliefs that are, in fact, false — or at any rate that are different from the beliefs of the imputing agent and that therefore will lead to actions that are different from those the imputing agent would have taken.

Studies of the development in children of this ability to impute mental states to others showed that by the age of 4 a normal child is aware that different people can have different beliefs about a situation [38]. By means of a test modelled on those by Wimmer and Perner in [38], namely the Sally-Anne experiment, it was shown that autistic children have a defective ToM by Baron-Cohen, Leslie, and Frith [6]. In the Sally-Anne experiment, children observe a scenario that can be enacted either by puppets or real people. The scenario involves two agents identified as Sally and Anne respectively. Sally plays with a toy (for example, a ball). She is seen to put the ball in a basket, and then to leave. In the absence of Sally, Anne is seen to move the ball to a different container, say a box. When Sally subsequently returns, the children are asked where Sally will look for the ball. To answer this question correctly, a child must realize that Sally has not seen the ball being moved and that Sally therefore believes, falsely, that the ball is still in the basket where she originally placed it. Autistic children consistently fail to answer correctly, and instead indicate that Sally will look for the ball where they, the autistic children, know it really is. One interpretation of this failure of mentalizing is to regard autistic children as possessing a rudimentary ToM in which the beliefs of other agents are assumed to be identical to those of the imputing agent. Evidence supporting this interpretation is that autistic children can form correct opinions of others' desires, and can act to sabotage the attempts by those other agents to achieve their goals, but cannot use deception for this purpose [26]. Thus autistic children are not entirely without a ToM, but suffer a specific deficit related to the representation of the (different) beliefs of others.

From the perspective of epistemic and doxastic logic, a psychologically realistic model of an agent capable of interacting intelligently in social situations would be one in which each agent a would possess a Theory of Mind comprising, relative to every agent b, a representation (which may or may not be accurate) of

b's state of mind (accessibility relation, preference relation). In the Sally-Anne experiment, the focus is not on the difference between weak and strong belief, but quite simply on the *presence* of belief. In experiments somewhat similar to the Sally-Anne experiment, observing children are not just asked about the actions of others but even about their beliefs, which comes even closer to true higher-order knowledge (an example is the 'Smarties' experiment in [5]).

A natural class of agent, suggested by the behaviour of *autistic* children, would consist of those agents a such that the ToM of agent a imputes to every agent b a state of mind (i.e. accessibility relation or preference relation) identical to a's own. The state of mind imputed by a to b is a surrogate, in a's own mind, of b's state of mind, and in this case the surrogate may be inaccurate, as b's accessibility relation (preference relation) may bear little resemblance to a's.[1] Three other natural classes of agent immediately leap to mind.

An *ideal*, as opposed to autistic, agent a would possess a ToM that imputes to every agent b a state of mind (accessibility relation, preference relation) identical to that actually possessed by b. Indeed, this is precisely the situation traditionally modelled by the possible worlds semantics in epistemic logic, whenever a statement such as "Agent a believes that agent b believes that p" implies that b believes p. On the level of semantics this means that the statement, which is primarily concerned with the beliefs of agent a, and which therefore is concerned with worlds accessible to agent a from the actual world, also in fact refers to the states accessible for agent b from the actual world, because of the accuracy of a's beliefs.

A *deranged* agent a has a ToM that imputes to agent b a state of mind that is wildly and randomly inaccurate, rather than suffering from the systematic inaccuracy of autistic agents. For example, patients with schizophrenia make false inferences about the intentions of others, such as assuming the intention to persecute [11].

The fourth and in some ways most interesting natural reference point for analysis is the *limited* agent a whose representation of b's state of mind is accurate as far as it goes but is incomplete. Such a limited agent differs from an autistic agent inasmuch as the latter makes an incorrect imputation of belief whereas a limited agent may simply refrain from imputing certain beliefs.

Consider, for example, that we are modelling an agent's state of mind by accessibility relations corresponding to preference relations that enable the agent to form both definite knowledge and defeasible beliefs. These two aspects may be considered separately, and we may therefore define a class of agent whose

[1] A logical alternative to assume a ToM deficiency in autistic agents is provided in [36, 29]. There, autism is seen as an executive dysfunction, which is operationalized by the autistic agent's inability to apply reasoning with exception handling, under circumstances where closed-world assumptions apply. This research is based on experimental results. There seems to be a strong link with the non-monotonic reasoning patterns that Stenning and Van Lambalgen observe in human agents, and the conditional / counterfactual reasoning based on partially ordered preferences, that we model explicitly in the logical language: having preferences means considering many exceptions, and ordering them. We also find it remarkable that they (properly) emphasize the executive aspect, but that their logical solution is not a dynamic logic.

members have partial representations of other agents' states of mind that are accurate with regard to knowledge but ignorant (and hence reticent) with regard to defeasible beliefs. In other words, we suggest that it is of interest to analyse the cognition of agent a where agent a has a ToM relative to each agent b that perfectly represents agent b's knowledge but that makes no imputation of defeasible beliefs (preferences) to agent b.

This is a theoretical definition, but it is not hard to conceive of circumstances in which it would be justified. A member of one culture encountering members of another culture may rationally assume that the strangers will observe the same gross facts about the world, but unless very naive will recognise that the strangers may make very different defeasible inferences to explain or extrapolate from such facts, and will recognise that ignorance is a poor basis for speculation about what such defeasible inferences may involve. One would therefore expect an anthropologist in the course of his professional duties to exemplify the class of limited agents. Another example, with a subtly different flavour, is that in which a human agent interacts with an artificial agent such as a database. If we assume that the entries in the database are carefully scrutinised for accuracy before and after insertion, so that responses to queries may be accorded the status of knowledge, and if the implementation makes no attempt to realise a closed-world assumption or other means of providing for queries to elicit defeasible inferences, the the human agent will adopt a ToM relative to the database which may accurately impute knowledge to the database but will make no imputation of defeasible beliefs, since the human agent knows that the database has no basis for the formation of such beliefs.

In the preceding two examples of limited agents, the limitation has been consciously self-imposed and has followed the line of cleavage between knowledge and defeasible belief. Limitations of ToM may follow other lines of cleavage, for example that separating the self from the rest of the environment. Agents may also have involuntarily limited ToM. By way of illustrating the combination of these features, consider that, although developmental studies of ToM have not been applied to psychopathic personalities, nevertheless there are some suggestive observations made by Cleckley:

> In a special sense the psychopath lacks insight to a degree seldom, if ever, found in any but the most seriously disturbed psychotic patients (...) He has absolutely no capacity to see himself as others see him. It is perhaps more accurate to say that he has no ability to know how others feel when they see him or to experience subjectively anything comparable about the situation. All the values, all of the major affect concerning his status, are unappreciated by him. [9, p.366]

In contrast to this limited blindness, psychopaths have sufficient insight into others' states of mind to be effective manipulators [14, pp.144–154].

Within the confines of our restricted modelling, this would suggest a class of agent in which agent a has an accurate representation of the beliefs and knowledge of agent b as these apply to matters other than agent a himself, and no representation of the beliefs of others as these may apply to him. In other words, there is accuracy up to a certain modal depth, and/or for specific stacks of modal operators, but not beyond. Such agents are therefore not fully introspective. They may, for example, only be introspective with regard to their beliefs about their own beliefs. In contrast, normal individuals have introspective facilities resembling their modal correspondents more closely.

It should be emphasised that ToM is not a philosophical or logical conjecture but a psychological faculty or module for which there is overwhelming empirical evidence. This evidence exists not only at the cognitive level of investigation, where tests such as the Sally-Anne experiment are applied, but also at the finer granularity of neuropsychology, where the discovery of mirror neurons have provided an explanation for meta-representation [12]. Cells have also been isolated in the prefrontal lobes which appear to be involved in the representation of the self [11].

3 Doxastic epistemic logic

The modal logical framework that we present as an example setting for knowledge and belief revision allows for revision of beliefs about other agents' beliefs. Degrees of belief are modelled by encoding so-called systems of spheres [18] in the structures on which these beliefs are interpreted. The systems of spheres formalize agent preferences. We call these semantic structures 'doxastic epistemic models' and they represent both degrees of belief and knowledge. They are multi-modal and multi-agent Kripke models. They are constructed from preference relations for agents, and for each agent they satisfy certain basic characterizable multi-agent frame properties. There are no multi-agent interaction axioms. With additional frame properties the structures interpret introspective belief and knowledge. Our presentation is a slight generalization of [34] (we assume partially ordered instead of totally ordered preferences), that was motivated by [10, 35], but as a framework it can be seen as 'logical folklore' and it is not dissimilar to the setup in [8, 1, 2].

Given are a set of atoms P, a set of agents A, and a partial order $\langle \mathcal{X}, < \rangle$ with a least element named 0.

Definition 1 (Doxastic epistemic model) A *doxastic epistemic model* is a triple $\langle S, <, V \rangle$. The set S is a *domain* of factual states, and *valuation* V is a function $V : P \to \mathcal{P}(S)$ such that each V_p is a subset of S. The *preference* function $<: A \to S \to \mathcal{P}(S \times S)$ defines a preference relation $<^s_a$ for each agent $a \in A$ and for each $s \in S$. The subset $\mathsf{domain}(<^s_a) \cup \mathsf{range}(<^s_a)$ of the domain S is the set $Plaus_a(s)$ of *plausible states* for agent a given s. There must be a *degree function* (as well written as) $<^s_a : Plaus_a(s) \to \mathcal{X}$ that is order preserving (i.e., if $t <^s_a t'$, then $<^s_a(t) < <^s_a(t')$) and that contains 0 as an image. A *doxastic epistemic state* is a pointed doxastic epistemic model, i.e., a structure $(\langle S, <, V \rangle, s)$ with $s \in S$. A doxastic epistemic *frame* is a doxastic epistemic model *without a valuation*, i.e., a pair $\langle S, < \rangle$. $\dashv$

We have abused the language and (also) write $<^s_a(s')$ for the degree/level of state s' in order $<^s_a$. The least degree of belief is 0. The degree function is not required to be surjective, e.g. (when $\mathcal{X}$ is a total order) $<^s_a(Plaus_a(s))$ may be an initial fragment of $\mathcal{X}$.

Definition 2 (Preference and degree of belief) Accessibility relation R^x_a is defined as: $R^x_a(s, s')$ iff $<^s_a(s') \leq x$; and $R^{\mathcal{X}}_a$ is defined as $\bigcup_{x \in \mathcal{X}} R^x_a$. $\dashv$

Definition 3 (Language of multi-agent doxastic epistemic logic)

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box^x_a\varphi \mid \Box^{\mathcal{X}}_a\varphi$$

Definition 4 (Semantics of doxastic epistemic logic) Let $\langle S, <, V\rangle$ be a doxastic epistemic model and $s \in S$. Then:

$$\begin{array}{lll} M, s \models \Box^x_a\varphi & \text{iff} & \text{for all } \ s' : R^x_a(s, s') \text{ implies } M, s' \models \varphi \\ M, s \models \Box^{\mathcal{X}}_a\varphi & \text{iff} & \text{for all } \ s' : R^{\mathcal{X}}_a(s, s') \text{ implies } M, s' \models \varphi \end{array}$$

For "$M, s \models \varphi$" read "formula φ is true in state s of model M." The truth of propositional connectives is defined as usual. A formula φ is *valid in a model* iff it is true in all states of that model, and a formula is *valid* iff it is valid in all models. A formula is *valid in a frame* iff it is valid in all models based on that frame. All these notions extend to formula schemata, as usual. From Definitions 1, 2, and 4 we immediately obtain the valid schemata, for each agent a:

- $\Box^x_a\varphi \rightarrow \Box^y_a\varphi$ iff $y \leq x$ *inclusion*
- $\Box^x_a\varphi \rightarrow \neg\Box^x_a\neg\varphi$ *seriality*

Principle $\Box^{\mathcal{X}}\varphi \rightarrow \Box^y\varphi$, for arbitrary $y \in \mathcal{X}$, follows easily from 'inclusion'; 'seriality' follows from $\Box^0_a\varphi \rightarrow \neg\Box^0_a\neg\varphi$ (seriality for normal belief) and 'inclusion'. With additional requirements: (let $x, y \in \mathcal{X}$ be arbitrary)

- $\Box^x_a\varphi \rightarrow \Box^y_a\Box^x\varphi$, and *arbitrary positive introspection*
- $\neg\Box^x_a\varphi \rightarrow \Box^y_a\neg\Box^x_a\varphi$, *arbitrary negative introspection*

the degrees of belief $\Box^x_a$ model (standard $KD45$) *introspective* degrees of belief and $\Box^{\mathcal{X}}_a$ models (introspective) *conviction* or 'strong belief'. If we demand as well that

- $\Box^{\mathcal{X}}_a\varphi \rightarrow \varphi$ *truth axiom*

then $\Box^{\mathcal{X}}_a$ models 'knowledge'. For degree 0 of introspective belief we write B_a instead of $\Box^0_a$, for degree $\mathcal{X}$ of introspective belief (conviction) we write $\boldsymbol{B}_a$ instead of $B^{\mathcal{X}}_a$, and if the truth axiom is also satisfied we write K_a instead of $\boldsymbol{B}_a$.

For models interpreting introspective belief, all states in the domain occupy a unique level in $\langle\mathcal{X}, <\rangle$; in other words, if $<^s_a(u) = x$ and $<^t_a(u) = y$, then $x = y$. We may then write $<_a(u)$ instead of $<^s_a(u)$. The set $Plaus_a(s)$ of *plausible states*

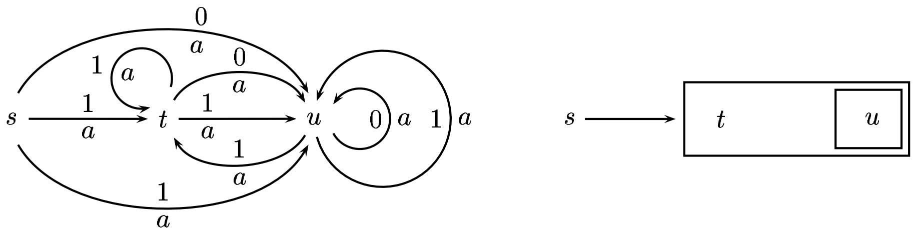


Figure 1: Two different visualizations of the same model. In the model on the left, all pairs in the accessibility relation are explicitly represented. Note that because of the inclusion relation $R_a^0 \subseteq R_a^1$, all 0-links are also 1-links. Also, transitivity is *not* assumed in the figure. In the model on the right, this is all implicit. State s is an inaccessible state pointing to the 'balloon' consisting of states t and u, such that u is preferred over t.

for agent a given s now corresponds to an $\to_a^{\mathcal{X}}$ epistemic class of the domain, and $<_a$ orders the states in that class. If all states are plausible, the domain is partitioned into such epistemic classes, in other words, it is partioned into disjoint orders $\langle \mathcal{X}, < \rangle$. The properties of knowledge ($K_a\varphi \to K_aK_a\varphi$, $\neg K_a\varphi \to K_a\neg K_a\varphi$), and properties relating belief to knowledge ($B_a\varphi \to K_aB_a\varphi$, $K_a\varphi \to B_a\varphi$) are derivable given these additional frame characteristics.[2]

Depicting multi-agent doxastic epistemic states We use two visualizations for Kripke models (see Figure 1). In the standard visualization an arrow between states s and t and labelled with a below and x above represents that $(s, t) \in R_a^x$. For multi-agent Kripke models for introspective agents, that makes quite a few arrows. In that case a simpler visualization is sufficient, namely one wherein for each agent and state, the set of all plausible states (i.e., the set of all states that are accessible from the given state, at any level x) can be visualized as a 'system of spheres' (or 'balloon'), with one sphere for each level x. All inaccessible states in the domain point to such a balloon. We have to do this for every agent. A state within some sphere relative to agent a, may be inaccessible relative to agent b. Therefore, states within balloons may point to, or belong to, other balloons, and so on, as in Figures 2, 3 and 4, later.

Example 5 Operators Anne (a) and Bill (b) have different access to the state of a fan and a light. Atomic proposition p stands for 'the fan is on' and atomic

[2] The 'base logic', without positive and negative introspection, and without the truth axiom, is complete. We only recently obtained that result and intend to report on this separately. It is not trivial, because 'individual knowledge' $\Box_a^{\mathcal{X}}$ is an *infinitary* modal operator in our approach, rather non-standardly. The completeness proof introduces *one* extra degree of belief. This is necessary to construct finite canonical models for formulas such as $\neg Kp \wedge Bp$. The logic remains complete (with respect to more restricted model classes) if one requires further axioms such as positive or negative introspection, or the truth axiom. This is because: these axioms have the Sahlqvist form [7, p.160], and are therefore frame-characterizable (see next Section), such that Kracht's Theorem can be applied [7, p.169].

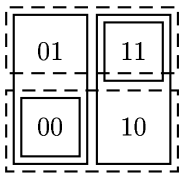


Figure 2: Two operators Anne and Bill have different access to the state of a fan and a light. Access for Anne is solid, access for Bill is dashed. Given that 00 is the actual state, the plausible states for Anne are 00 and 01. But from the perspective of state 01, the plausible states *for Bill* are 01 and 11. Therefore, the epistemic classes for Anne and Bill overlap.

proposition q stands for 'the light is on'. The doxastic epistemic model in Figure 2 represents a situation where Anne knows whether the fan is on, whereas Bill knows whether the light is on. Also, if Anne knows that the fan is on, she believes that the light is on, and if she knows that the fan is off, she believes that the light is off. Note that in Figure 2, access for a is drawn with solid lines and access for b is drawn with dashed lines.

For both agents, access consists of two equivalence classes, and there are no implausible states. Anne's preference among the states in her equivalence classes is that 00 $<_a$ 01 and 11 $<_a$ 10. Therefore, the state named 00 is in a square that is contained in the square also containing 01. Anne's preference does not depend on which is the actual state in the class, so that we can drop the perspective of the actual state in our notation: we write 00 $<_a$ 01 because both 00 $<_a^{00}$ 01 and 00 $<_a^{01}$ 01. Anne's equivalence class $\{00, 01\}$ represents that $\{(01, 00), (00, 00)\} = R_a^0$ and that $\{(01, 00), (00, 00), (00, 01), (01, 01)\} \in R_a^1$ ($R_a^0 \subseteq R_a^1$). In this case, $R_a^{\{0,1\}} = R_a^1$. Bill has no preferences among plausible states, therefore $R_b^0 = R_b^1$ ($= R_b^{\{0,1\}}$).

We can now evaluate statements of knowledge and belief. In the state where the fan and light are both off, Anne believes that, and believes that Bill knows that the light is off. In the state where the fan is off and the light is on, Anne knows that the fan is off, and she (incorrectly) believes that the light is off; she even believes that Bill knows that the light is off, even though Bill actually knows that the light is on. Formally—let M'' be the model in Figure 2:

$$\begin{aligned} &M'', 00 \models B_a(\neg p \wedge \neg q) \wedge B_a K_b \neg q \\ &M'', 01 \models \neg p \wedge K_a \neg p \wedge B_a \neg q \wedge B_a K_b \neg q \wedge K_b q \end{aligned}$$

For example, $M'', 00 \models B_a(\neg p \wedge \neg q)$ because $M'', s \models \neg p \wedge \neg q$ in all states s that are R_a^0-accessible from 00. As this is 00 only, we check whether $M'', 00 \models \neg p \wedge \neg q$. This is true because $00 \notin V_p$ and $00 \notin V_q$. ⊣

4 Case studies in belief interaction

In the following subsections we model that agents believe that other agents do not distinguish among their beliefs ('no preferences'), that other agents have the same beliefs ('my preferences'), and that other agents have *some* of the same beliefs ('preference refinement'). The final subsection tentatively presents other common preference dependencies.

All these multi-agent belief dependencies are frame characterizable. A formula schema φ is *frame characterizable* when it is valid on a frame if and only if the frame satisfies a certain first-order definable property. For example, the validity $\Box^x_a \varphi \to \Box^y_a \Box^x \varphi$ for arbitrary positive introspection (Section 3) corresponds to the frame property "for all $s, t, u \in S$, if $R^y_a(s,t)$ and $R^x_a(t,u)$, then $R^x_a(s,u)$."

The correspondence results in the coming sections are elementary. They follow from the Sahlqvist form of the axioms [7, p.160]. From a *technical* logical point of view, the issue therefore lacks interest. But we think that in view of modelling multi-agent systems, the matter is very interesting indeed: multi-agent interaction axioms occur rarely in epistemic logics[3].

Model characterizable agent interaction may have received more attention than frame characterizable agent interaction, see e.g. the characteristic formulas describing pointed models in [30]. But the situation-independent agent stance in ToM seems to be more suitably modelled with frame characteristics as these are valuation independent, i.e., the agent interaction is not formulated in terms of actual situations.

4.1 No preferences

In the absence of information, a wise stance may be to reason from your own preferences, but to assume that other agents have no preferences. A further restriction that one can make is that you not only 'normally believe' other agents to have no preferences, but are also convinced of (strongly believe) that. And an even further restriction is that you assume this conviction in others as well, and that they have this conviction about your conviction, etc. ad infinitum. This is formalized in the characterizing axiom

$$\Box^x_a (\Box^y_b \varphi \to \Box^z_b \varphi) \qquad \textit{(you have) no preferences}$$

where all of x, y, z are arbitrary. Because y and z are arbitrary, this entails that for arbitrary $y \in \mathcal{X}$, $\Box^x_a (\Box^y_b \varphi \leftrightarrow \Box^{\mathcal{X}}_b \varphi)$. This says that agent a has an arbitrarily high degree of confidence (is convinced) in arbitrary beliefs of others to correspond to their knowledge/conviction. In particular, normal beliefs of others correspond to their convictions. Therefore, all plausible states are most (and therefore equally) preferred: the other agent has no preferences.

[3] We only know of a property called 'weakly directedness' which characterizes the dependencies in hypercubes [20]. A hypercube is the Kripke model equivalent of a kind of interpreted system (namely the full cartesian product of local state value sets), a well-known abstract architecture for multi-agent systems.

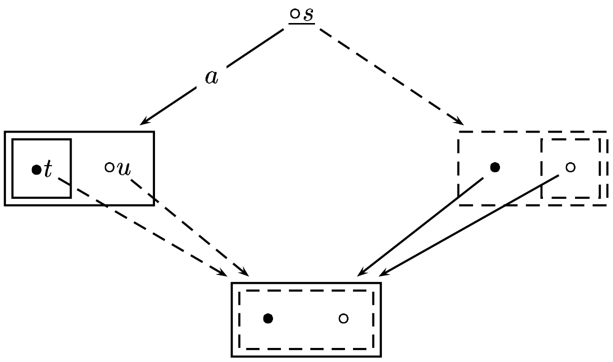


Figure 3: Minimal doxastic epistemic state wherein agents assume that other agents have no preferences. Only some states have been named. Atom p is true in • states, and false in ∘ states. The actual state s is underlined. It is not plausible for either agent. Given s, the plausible states for a are (the system of two spheres for states) t and u, where t is preferred over u. This means that $R_a^0 = \{(s,t),(t,t),(u,t)\}$ and that $R_a^1 = \{(s,t),(t,t),(u,t),(s,u),(t,u),(u,u)\}$. Given state t the plausible states for agent b are the bottom two in the model, where b has no preference about p; and similarly for u. Therefore, in actual state s agent a believes that agent b has no preferences.

Proposition 6 $\Box_a^x(\Box_b^y\varphi \to \Box_b^z\varphi)$ corresponds to frame property "for all s,t,u, if $R_a^x(s,t)$ and $R_b^z(t,u)$, then $R_b^y(t,u)$." ⊣

Proof The proof is elementary. We present it as an example. Other proofs of correspondence results are omitted.

⇒ Suppose the property does not hold in some model M. Then there are s,t,u in its domain and $x,y,z \in \mathcal{X}$ such that $R_a^x(s,t)$ and $R_b^z(t,u)$ *but not* $R_b^y(t,u)$. Note that this implies that $y < z$ (if, hypothetically, $z \leq y$, then $R_a^z \subseteq R_a^y$, which contradicts the assumption). Now let some atom p be true everywhere in M except in u. Then we have that $M,t \models \Box^y p$ but not $M,t \models \Box^z p$, so that $M,s \not\models \Box_a^x(\Box_b^y p \to \Box_b^z p)$.

⇐ ('Trivial') Suppose the property holds. Let model M and state s be arbitrary. To show that $\Box_a^x(\Box_b^y\varphi \to \Box_b^z\varphi)$ is true in s for all φ, assume an arbitrary t with $R_a^x(s,t)$. We then have to show that $\Box_b^y\varphi \to \Box_b^z\varphi$ is true in t. To show that, assume (take the dual form) an arbitrary u such that $R_b^z(t,u)$. We then have to show that $R_b^y(t,u)$. This follows directly from applying the property. □

Figure 3 depicts a model satisfying the property, for introspective agents. The model is the smallest that satisfies the property and in which two agents have different, incorrect, beliefs. Some truths and validities in this model M are (note that $\Box_a^0 = B_a$ and $\Box_a^1 = \Box_a^{\{0,1\}} = \boldsymbol{B}_a$, and similarly for b):

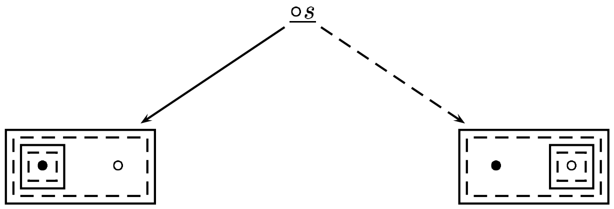


Figure 4: Smallest, five-state, doxastic epistemic state wherein both agents are incorrectly convinced that the other agent has the same preferences. Atom p is true in • states, false in ∘ states.

$$\begin{array}{l} M, s \models \neg p \\ M, s \models B_a p \\ M, s \models B_b \neg p \\ M, s \models B_a \neg B_b \neg p \\ M, s \models \boldsymbol{B}_a \neg B_b \neg p \\ \models \boldsymbol{B}_a (B_b \varphi \leftrightarrow \boldsymbol{B}_b \varphi) \end{array}$$

4.2 My preferences

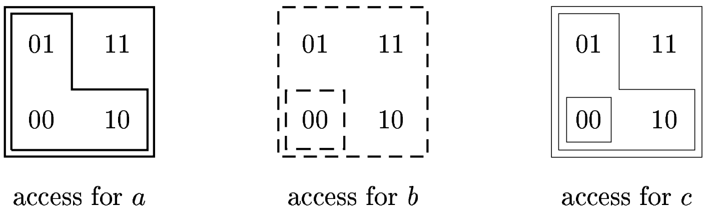
In this case, all agents assume that other agents have the same beliefs as themselves, regardless of these agents' actual preferences (and, again, everyone also assumes everyone else to be like that, etc.). The extreme case is the autistic agent in Section 2. (It would be problematic to explain how such a society of autistic agents came about, i.e., what form of belief revision could would result in such a situation: see Section 5 on belief revision.) In a slightly different modal setting (namely where 'preferred' means 'more desirable') this is quite typical for people buying birthday presents for their friends and family: it is very hard to buy a present for your friend unless it also suits your own tastes. The characterizing axiom is

$$\Box^x_a(\Box^y_a \varphi \leftrightarrow \Box^y_b \varphi) \qquad \textit{(you have) my preferences}$$

Proposition 7 $\Box^x_a(\Box^y_a \varphi \leftrightarrow \Box^y_b \varphi)$ corresponds to frame property "for all s, t, u, if $R^x_a(s, t)$, then $R^y_a(t, u)$ iff $R^y_b(t, u)$." ⊣

A minimal model M wherein agents incorrectly believe that others have the same preferences is depicted in Figure 4. In that figure we have that:

$$\begin{array}{l} M, s \models \neg p \\ M, s \models B_a p \\ M, s \models B_a B_b p \\ M, s \models B_b \neg p \\ M, s \models B_b B_a \neg p \\ \models \boldsymbol{B}_a (B_b \varphi \leftrightarrow B_a \varphi) \end{array}$$



access for a access for b access for c

Figure 5: We depict three different perspectives, namely for agents a, b, and c of a (one and the same) four-state model. (To avoid confusion the different accessibility relations are not visually superimposed onto the same model, as in previous figures.) Access for c refines access for a, and for b, but access for b does *not* refine access for a.

4.3 Preference refinement

We now model that an agent believes that other agents' beliefs are 'somewhat' like his own—in the sense that he believes that his own preference relation is a *refinement* of other's. A special case is where he believes other agents' preferences to be *exactly* as his own, as in Subsection 4.2. And another special case is that of interpreted systems (or other multi-agent $S5$ models) where *some* agents **know** that their preferences are a refinement of those of other agents (this case is not interesting if *all* agents know that, because all preferences are then equal).

How to formalize refinement between preference orders? A naive and incorrect way to do that, is to ensure that corresponding accessibility relations for all degrees of preference satisfy inclusion. Given two agents a and b, axiom schema $\Box_a\varphi \rightarrow \Box_b\varphi$ describes that b knows (or believes) more than a, which corresponds to $R_b \subseteq R_a$. The generalization of that to sets of modal operators satisfying inclusion, is "for all $x \in \mathcal{X}$, $\Box^x_a\varphi \rightarrow \Box^x_b\varphi$." Example 8 shows that this approach does *not* work.

Example 8 See Figure 5. Consider two atoms p and q and states named 00, 01, 10, and 11—where 01 is the state such that p is false and q true, and three orders $00 =_a 01 =_a 10 <_a 11$ (for agent a), $00 <_b 01 =_b 10 =_b 11$ (for agent b) and $00 <_c 01 =_c 10 <_c 11$ (for agent c). Both $B_a\varphi \rightarrow B_b\varphi$ and $\boldsymbol{B}_a\varphi \rightarrow \boldsymbol{B}_b\varphi$ are valid on the model; as we have that $R^0_b = \{(00, 00)\}$ and $R^0_a = \{(00, 00), (01, 00), \dots\}$, so that $R^0_b \subseteq R^0_a$, and both R^1_a and R^1_b are the universal relation. Relation $<_b$ does not refine $<_a$, because agent a can distinguish between 10 and 11, but agent b cannot. Relation $<_c$ is a refinement of $<_a$, and also of $<_b$, in our intended meaning. $\dashv$

There is a proper way to formalize refinement if order $\langle \mathcal{X}, < \rangle$ is a well-order. For $<^s_a$ to be a *refinement* of $<^s_b$, we require that $t <^s_b t'$ implies $t <^s_a t'$, and each degree y in $<^s_b$ corresponds to a *set* of degrees x in $<^s_a$ that is *dense* in $\mathcal{X}$ (i.e., if x', x'' in that set, and $x' < x''' < x''$, than x''' is also in that set). This induces an equivalence $\sim$ on degrees x in $<^s_a$. Therefore we may write $[x]$ for levels y in $<^s_b$, where $[x] = \{x' \in \mathit{Plaus}_a(s) \mid x' \sim x\}$. As usual, one then defines sets of modal operators $\Box^x_a$ and $\Box^{[x]}_b$. The characterizing axiom expressing that (each) a believes that b's preferences are part of his own, i.e., that a's preference system is a refinement of b's, is

$$\Box^y_a(\Box^{[x]}_b\varphi \to \Box^x_a\varphi) \qquad\qquad \textit{preference refinement}$$

Proposition 9 $\Box^y_a(\Box^{[x]}_b\varphi \to \Box^x_a\varphi)$ corresponds to frame property "for all s, t, u, if $R^y_a(s,t)$ and $R^x_a(t,u)$, then $R^{[x]}_b(t,u)$." $\dashv$

For the special case of knowledge, such that all beliefs are truthful, we get something slightly different. Obviously, the only case where *all* agents *correctly* believe their preferences to be a refinement of those of others is when all preferences are equal for all agents: you correctly believe that about someone else, but (s)he also *correctly* believes that about you! The different case where one preference system is a refinement of another comes with the correspondence:

Corollary 10 $\Box^{[x]}_b\varphi \to \Box^x_a\varphi$ corresponds to frame property "for all s, t, if $R^x_a(s,t)$, then $R^{[x]}_b(s,t)$." $\dashv$

Example 11 In Example 8, $<_c$ is a refinement of both $<_a$ and $<_b$ according to this interpretation. Consider the first. Degrees 0 and 1 in $<_c$ correspond to degree 0 in $<_a$, and degree 2 in $<_c$ corresponds to degree 1 in $<_a$. As $\{0,1\}$ is an initial segment of $\{0,1,2\}$, we further have that $R^2_a = R^1_a$. We now have that $B_a\varphi \to B^1_c\varphi$ (we write B^1_c for $\Box^1_c$) and that $\boldsymbol{B}_a\varphi \to \boldsymbol{B}_c\varphi$. As $B^1_c\varphi \to B_c\varphi$, we then also have that $B_a\varphi \to B_c\varphi$. $\dashv$

4.4 Other preference interaction

The three ways to model preference interaction are just examples of such 'agent stance', as backed up by 'theory of mind' evidence concerning autism, psychopaths, and artificial agents. Many other settings are conceivable, that may have similar correspondents in psychology and that may realistically model common patterns of preference interaction.

To assume that you are *convinced* that others have no preferences is 'a bit much'. The assumption that others have no preferences is made given the absence of any information, or out of politeness. Even if you know your neighbour to be violently opposed to cutting down that tree, the convential way to start a discussion on that matter is to ask him whether he has an opinion on the matter—i.e., you 'present' him with a model wherein he is supposed to believe that you believe that he has no preference. That usually does not describe strong but only tentative belief. After the neighbour has shouted you down, as usual, there is no point in keeping up appearances, and (given the current perspective of the neighbour) you revise your tentative belief that he has no preferences into one where he has. That revision is not 'surprising' or endangering any convictions; nor is that revision resulting in inconsistency—even though it drives you mad indeed. The setting wherein you (a) only *tentatively* believe that others (b) have no preferences has characterizing axiom

$$\Box^0_a(\Box^y_b\varphi \to \Box^z_b\varphi)$$

i.e., in the form of a corollary for introspective agents:

$$B_a(B_b\varphi \to \boldsymbol{B}_b\varphi)$$

This expresses that agent a (tentatively) believes that agent b is convinced of all his beliefs. This is also frame characterizable.

There are also common patterns of communication wherein the default is to assume that the beliefs or knowledge of the other are the *opposite* of your own. Consider a stranger approaching you on the street and asking you: "Do you know the way to the railway station?" This tourist lost in town who does not know the way, tentatively assumes that you know the way. In other words: in the absence of other information, the assumption is that another agent's knowledge (as an extreme case of preference) is different from its own.[4]

We have not given settings where a large number of degrees of belief is essential. Our examples involved at most three degrees of belief. The main reason for that was to keep the exposition simple. Typical scenarios wherein more degrees of belief play a role involve thresholds needed for action; an agent collects evidence thus gradually increasing his degree of belief or confidence in a certain possible state of affairs, and only acts after a threshold of confidence has been exceeded. For another example, the interval $[0,1]$ is an order $\mathcal{X}$ allowed by our framework (see Definition 1)—this suggests certain correspondences between reasoning with plausibilities and reasoning with possibilities, but to explain that would carry too far [13, 34].

The evidence for non-introspective agents [14] as discussed in Section 2 suggested 'restricted modal depth' ToM. This may point in the direction of yet other frame-characterizable properties of multi-agent belief interaction. Obviously, real agents are never fully introspective for computational reasons, but in non-psychopatic everyday communication the assumption frequently is that agents are—for reasons of either politeness, (mental) computational efficiency, or (as in [17]) convention.

We repeatedly referred to belief and knowledge *revision* subject to additional information. A valid question is which frame characteristics and corresponding axioms are preserved under belief revision. This will now be addressed.

5 Belief revision

5.1 Dynamic operators for belief change

To the doxastic epistemic logic defined in Section 3 we can add a dynamic modal operator to express various forms of belief revision. One adds one more inductive clause to the language definition (Definition 3), namely $[\circledast\varphi]\psi$. As in Section 3 we follow the approach in [34]. It is not unlike that in **DDL** [25, 19]. Expression $[\circledast\varphi]\psi$ stands for 'after belief revision with φ, ψ holds'. The operator $[\circledast\varphi]$ is a *dynamic* modal operator. It is interpreted as a doxastic epistemic state

[4] This example originates with Johan van Benthem, in a slightly simpler setting.

transformer, i.e., as some sort of ‘program’. Its semantics is defined as

$$M, s \models [\circledast\varphi]\psi \text{ iff for all } (M', s') : (M, s)[\![\circledast\varphi]\!](M', s') \text{ implies } M', s' \models \psi$$

where (M', s') is a transformed doxastic epistemic state, according to some computation relative to (M, s) and the revision formula φ. There are only two obvious ways in which such models can be transformed: the set of plausible states relative to a given state in the domain can be reduced, *or* the ordering between states in that set of plausible states can be changed. Which mechanism is chosen depends on one’s preferred belief revision mechanism. Various ways are distinguished in [27, 2, 34, 24, 31, 4]. Note that in this dynamic setting, belief revision can be iterated as well—Rott even distinguishes 27 ways to do that in [24]. We outline three ways of ‘dynamic modal’ belief change.

Belief expansion as public announcement In the case of *public announcement*, for each state in the domain we restrict the set of plausible states to those where the announced formula is true, and keep the order between the remaining plausible states. This can be accomplished by restricting the domain of model M' above to those states where the revision formula is true, and readjusting the levels of those remaining states. This adjustment is necessary, because if the most plausible (preferred) states—with level 0—were removed by the process of domain restriction, other states now have become most plausible and thus should get level 0. The construction can be seen as an implementation of belief *expansion*, and is a mere adjustment of the public announcements in [21, 3]. This setting might best be seen as one that models change of knowledge only—as a result of incoming definite information—any change of tentative beliefs is a function of that knowledge change. It is somewhat related to the proposal by Van Benthem and Liu in [32] wherein there also is no interaction between upgrade (i.e., preference change) and update.

Minimal-Spohn belief revision In the case of belief revision coined by Rott as ‘minimal-Spohn’ [27, 24], in the computational setting suggested by Aucher [2], the set of plausible states for every state remains the same, but their ranking may be changed. This is accomplished by changing their ranking relative to the revision formula (seen as a ‘tentative public announcement’). To keep things simple, assume introspective agents so that the ranking is independent from the perspective of the actual state. Let M be the model before revision, then the model M' resulting from revision is as M except with belief function $<'$ defined as (see [2, 34]):

$$\begin{array}{lcll} <'(s) & = & <(s) - \mathsf{Min}\{<(t) \mid M, t \models \varphi\} & \text{if } M, s \models \varphi; \\ <'(s) & = & <(s) + 1 - \mathsf{Min}\{<(t) \mid M, t \models \neg\varphi\} & \text{otherwise} \end{array}$$

In other words: the most normal states among the φ-states become the most normal states in the model (by ensuring that their level becomes 0), whereas the most normal states among the $\neg\varphi$-states become *definitely* no longer the most normal states in the model (by ensuring that their level becomes 1). (A detail corresponding to removing possible ‘gaps’ between degrees has been omitted from this presentation.)

Maximal-Spohn belief revision An interesting related revision operator, also distinguished in [27, 24], is known as *maximal Spohn.*[5] Just as in minimal Spohn, the relative order among φ-states, and among $\neg\varphi$-states, is preserved, but, unlike there, *all* $\neg\varphi$-states have become less preferred than *all* the φ-states. This can be implemented by changing the 'otherwise' clause for minimal-Spohn into

$$<'(s) \quad = \quad <(s) + 1 - \mathsf{Min}\{<(t) \mid M, t \models \neg\varphi\} + \mathsf{Max}\{<(t) \mid M, t \models \varphi\}$$

5.2 Preservation of belief interaction under belief change

We can now investigate which of the three agent interaction axioms 'no preferences', 'my preferences', and 'preference refinement', as distinguished in the previous section, are preserved under these three belief change operations. Most of the following results are obvious and therefore go without proof.

Proposition 12 'No preferences', 'my preferences', and 'preference refinement' are preserved under belief expansion. ⊣

Proof All three interaction axioms are universally quantified expressions and therefore preserved under submodel restriction. Belief expansion is a submodel restriction, and submodel restrictions are order preserving between states. □

This corresponds to our intuition that change of *knowledge* should not interact with distinctions between *degrees of belief*.

For the essentially preference-adjusting minimal-Spohn and maximal-Spohn revision we do not necessarily expect such results. Yet, again, there are surprising results that may be considered interesting.

Proposition 13 'No preferences' is not preserved under minimal-Spohn revision and also not preserved under maximal-Spohn revision. ⊣

A counterexample for both forms of belief revision is provided in Example 14. This result is according to our intuition that assumed tentative beliefs in others (such as the absence of belief distinctions) can be updated by communicative interactions or shared observation.

Example 14 If I believe that you have no preferences about atom p (or some other boolean expression—for modal formulas the situation is more complex), and you believe that about me, and we both revise with p (in this multi-agent setting the revisions modelled all represent *public* forms of belief revision), then I should afterwards believe that you prefer p over $\neg p$—like I do. Therefore, the frame correspondence described by me believing that you have no preferences, is *not* preserved under belief revision. See the bottom state transition in Figure 6 for an example—this transforms the doxastic epistemic state previously depicted in and discussed ad Figure 3. ⊣

[5] Recent contributions to the literature name this form of revision 'lexicographic upgrade' [31] and 'anti-lexicographic update' [4].

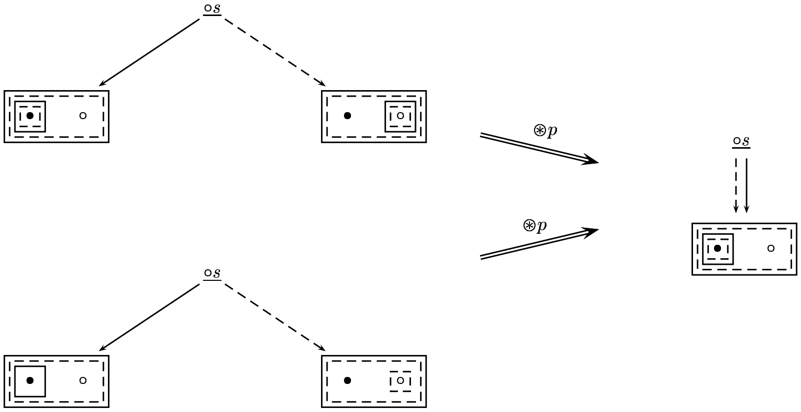


Figure 6: My belief that your preferences are as my own, is preserved under belief revision (top transition), but my belief that you have no preferences, is not preserved under belief revision (bottom revision).

Proposition 15 'My preferences' is preserved under minimal-Spohn and under maximal-Spohn revision. ⊣

Proof If I am *convinced* that your beliefs are as mine, and if I also *know* that you update with the same revision formula as I do, then I will afterwards still be convinced that your beliefs are as mine. □

If I believe that your preferences are exactly as mine, then any revision will *still* satisfy that description—my preferences may change, but, identically, so will yours in that case. Also, if our actual preferences were different before the revision, they may now be the same. The top state transition in Figure 6 provides an example (see Figure 4). Proposition 15 *may* be construed as evidence for the 'closed world' in which autistic agents appear to live: they cannot change their belief in other agents having their beliefs. But that would improper: more complex forms of belief revision [34, 4] are needed to model autistic agents' revision strategies (namely strategies where the autistic individual's revision is private and differs from other simultaneously present individuals' revision) than the public forms of belief revision that we distinguish here, in which the agents have common knowledge of the nature of the change. Only such more complex mechanisms can describe that a very young child's belief *becomes* different from (theoretically) other present children's beliefs in the position of the ball after initially having been the same.

Proposition 16 'Preference refinement' is not preserved under minimal-Spohn revision. ⊣

Two states that are equally preferred in the 'less refined system' but distinguishable in the 'more refined system' may after minimal-Spohn revision be distinguishable in the 'less refined' system but equally preferred in the 'more refined' system. This is due to the peculiarity of minimal-Spohn revision that the $\neg\varphi$-states are 'renormalized' to a, possible low, degree of preference. This may interact with preferences among φ-states. See Example 17.

Example 17 Consider two agents a and b, and five states c, d, e, f, g described by atomic propositions of the same name that are only true in the state with that name, and that are ordered as

$$c <_a d <_a e =_a f <_a g$$
$$c =_b d =_b e =_b f <_b g$$

for agent a and agent b, respectively. In other words, agent a considers state c most likely, and distinguishes four degrees of belief, whereas agent b considers the states c, d, e most likely, etc; $<_a$ is a refinement of $<_b$. We also assume that for both agents the preference distinctions are 'just those needed' in the sense that there are only four degrees of belief *anyway*, i.e., $<_a(c) = 0, <_a(d) = 1, <_a(e) = <_a(f) = 2$ and $<_a(g) = 3$ and analogously for agent b. Now execute minimal-Spohn belief revision on this two-agent system for formula $c \vee d \vee e$ (apply the definition on page 16). The outcome is the doxastic epistemic model with preferences

$$c <_a d =_a f <_a e =_a g$$
$$c =_b d =_b e <_b f <_b g$$

After the revision, $<_b$ no longer refines $<_a$, because agent a has become indifferent between states d and f, unlike agent b who now distinguished d from f. ⊣

Unlike minimal-Spohn revision, maximal-Spohn revision retains all the original preference distinctions (obviously—it is a refinement operation plus reshuffling). With that nice result we close the section.

Proposition 18 'Preference refinement' is preserved under maximal-Spohn revision. ⊣

6 Further research

Our results were not restricted to introspective agents, but our examples were. We would like to pursue modelling various forms of non-introspective multi-agent interaction, also with regard to the computational limitations of agents inasfar as that is compatible with the requirements of frame characterizability.

The belief revision procedures presented were all public forms of belief revision: when presented with new information, all agents react to it in the same way. Clearly, psychologically realistic settings for belief revision also involve more complex forms, e.g. including 'private' belief revision of some individuals in a larger group (see the observations on autism ad Proposition 15 on page 18). There are various available logics to model that [2, 34, 4].

We closed the previous section on belief revision with various results on the relation between multi-agent interaction and belief revision strategies. This may be a fruitful path to pursue further: one might venture to suggest that cognitively realistic patterns of multi-agent interaction should be preservable under belief revision. In other words, one's preferred belief revision mechanism restricts what forms of (static) multi-agent interaction are acceptable: interaction that is preserved under belief revision may be preferable. That would be one more way in which psychology could affect logical modelling.

But there are also logical reasons to prefer multi-agent interaction that is preserved under belief revision: the corresponding dynamic epistemic logics would be complete for certain multi-agent classes of models (see footnote 2—this applies just as well to any frame characterizable axiom). If anything, that suggests a direction in which logic might affect psychology: computational features of logics may be relevant for ToM, by providing 'computationally cheap' explanations for observed behaviour.

Also linked to computational efficiency, and the direction from logic to psychology, is that acquiring 'theory of mind' can be seen as learning social conventions. Conventions are a form of common knowledge (or 'background knowledge'). Higher-order preferences can be 'enforced' by the *assumption* of common knowledge of learning, even though such common knowledge may not practically be obtainable (or even intractible: fixed-points take time to compute, and the structures may be large or infinite). Such an approach would require a new logic to match it. A convincing dynamic logical model for 'jumping to common knowledge' might also in due time provide a new insight to the theory of mind.

References

[1] G.B. Asheim and Y. Søvik. Preference-based belief operators. *Mathematical Social Sciences*, 50(1):61–82, 2005.

[2] G. Aucher. A combined system for update logic and belief revision. Master's thesis, ILLC, University of Amsterdam, Amsterdam, the Netherlands, 2003.

[3] A. Baltag and L.S. Moss. Logics for epistemic programs. *Synthese*, 139:165–224, 2004. Knowledge, Rationality & Action 1–60.

[4] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. Proceedings of LOFT 2006 (7th Conference on Logic and the Foundations of Game and Decision Theory), 2006.

[5] S. Baron-Cohen. *Mindblindness: An essay on autism and theory of mind.* MIT Press, Cambridge, MA, 1995.

[6] S. Baron-Cohen, A.M. Leslie, and U. Frith. Does the autistic child have a 'theory of mind'? *Cognition*, 21:37–46, 1985.

[7] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic.* Cambridge University Press, Cambridge, 2001. Cambridge Tracts in Theoretical Computer Science 53.

[8] O. Board. Dynamic interactive epistemology. *Games and Economic Behaviour*, 49:49–80, 2004.

[9] H. Cleckley. *The mask of sanity.* Mosby, St Louis, MO, 1976. Available as `http://www.cassiopaea.org/cass/sanity_1.pdf`.

[10] D. Ferguson and W.A. Labuschagne. Information-theoretic semantics for epistemic logic. In *Proceedings of LOFT 5*, Turin, Italy, 2002. ICER. no page numbers.

[11] C.D. Frith and U. Frith. Interacting minds — a biological basis. *Science*, 286:1692–1695, 1999.

[12] V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 12:493–501, 1998.

[13] J.Y. Halpern. *Reasoning about Uncertainty.* MIT Press, Cambridge MA, 2003.

[14] R.D. Hare. *Without conscience.* The Guilford Press, New York, 1993.

[15] S. Kraus and D. Lehmann. Knowledge, belief and time. *Theoretical Computer Science*, 58, 1988.

[16] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[17] D.K. Lewis. *Convention, a Philosophical Study.* Harvard University Press, Cambridge (MA), 1969.

[18] D.K. Lewis. *Counterfactuals.* Harvard University Press, Cambridge (MA), 1973.

[19] S. Lindström and W. Rabinowicz. DDL unlimited: dynamic doxastic logic for introspective agents. *Erkenntnis*, 50:353–385, 1999.

[20] A.R. Lomuscio. *Knowledge Sharing among Ideal Agents.* PhD thesis, University of Birmingham, Birmingham, UK, 1999.

[21] J.A. Plaza. Logics of public communications. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, and Z.W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

[22] D. Premack and G. Woodruff. Does the chimpanzee have a 'theory of mind'? *Behavioral and Brain Sciences*, 4:515–526, 1978.

[23] Anand S. Rao and Michael P. Georgeff. Modeling rational agents within a BDI-architecture. In James Allen, Richard Fikes, and Erik Sandewall, editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484, San Mateo (CA), 1991. Morgan Kaufmann.

[24] H. Rott. Adjusting priorities: Simple representations for 27 iterated theory change operators. Manuscript, 2004.

[25] K. Segerberg. Two traditions in the logic of belief: bringing them together. In H.J. Ohlbach and U. Reyle, editors, *Logic, Language, and Reasoning*, pages 135–147, Dordrecht, 1999. Kluwer Academic Publishers.

[26] B. Sodian and U. Frith. Deception and sabotage in autistic, retarded, and normal children. *Journal of Child Psychology and Psychiatry*, 33:591–606, 1992.

[27] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume II, pages 105–134, 1988.

[28] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

[29] K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. MIT University Press, Cambridge MA, 2006.

[30] J.F.A.K. van Benthem. Dynamic odds and ends. Technical report, ILLC, University of Amsterdam, 1998. Report ML-1998-08.

[31] J.F.A.K. van Benthem. Dynamic logic for belief change. *Journal of Applied Non-Classical Logics*, 2006. To appear.

[32] J.F.A.K. van Benthem and F. Liu. Dynamic logic of preference upgrade. Technical report, University of Amsterdam, 2005. Report PP-2005-29.

[33] W. van der Hoek. Systems for knowledge and beliefs. *Journal of Logic and Computation*, 3(2):173–195, 1993.

[34] H.P. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese (Knowledge, Rationality & Action)*, 147:229–275, 2005.

[35] H.P. van Ditmarsch and W.A. Labuschagne. A multimodal language for revising defeasible beliefs. In E. Álvarez, R. Bosch, and L. Villamil, editors, *Proceedings of the 12th International Congress of Logic, Methodology, and Philosophy of Science (LMPS)*, pages 140–141. Oviedo University Press, 2003.

[36] M. van Lambalgen and H. Smid. Reasoning patterns in autism: rules and exceptions. In *Proceedings of the Eighth International Colloquium on Cognitive Science*. Kluwer Science Publishers, 2003.

[37] F.P.J.M. Voorbraak. *As Far as I Know.* PhD thesis, Utrecht University, Utrecht, NL, 1993. Questiones Infinitae volume VII.

[38] H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128, 1983.