A Logic for Emotions: a basis for reasoning about commonsense psychological knowledge

Kathryn E. Sanders
Department of Computer Science, Brown University

Abstract

There is a body of commonsense knowledge about human psychology that we all draw upon in everyday life to interpret our own actions and those of the people around us. In this paper, we define a logic in which this knowledge can be expressed. We focus on a cluster of emotions, including approval, disapproval, guilt, and anger, most of which involve some sort of ethical evaluation of the action that triggers them. As a result, we are able to draw on well-studied concepts from deontic logic, such as obligation, prohibition, and permission. We formalize a portion of commonsense psychology and show how a simple problem can be solved using our logic.

1 Introduction

There is a body of commonsense knowledge about human psychology that we all draw upon in everyday life to interpret our own actions and those of the people around us. This knowledge is brought to bear by the reader in interpreting texts involving human actions and reactions.

In this paper, we define a logic in which commonsense knowledge about human psychology can be expressed. We formalize a portion of this knowledge and show how a simple problem can be formulated and solved using this logic. We focus on a cluster of emotions, including approval, disapproval, guilt, and anger, most of which involve some sort of ethical evaluation of the action that triggers them. As a result, we are able to draw on well-studied concepts from deontic logic, such as obligation, prohibition, and permission, in formalizing this knowledge.

In order to handle concrete problems, since emotions do not occur in a vacuum, it is also necessary to formalize some commonsense knowledge about actions and the probable evaluation of those actions by the agents and others. Specifically, we focus on a cluster of actions having to do with ownership and possession of property – giving, lending, selling, and stealing – and the predictable responses to those actions. We demonstrate that our logic is sufficiently expressive to handle a variety of information about human actions and responses, in a way that is substantially more formal than previous work in this area.

In this paper, we focus on the following example:

Jack went to the supermarket. He parked his car in a legal parking place. When he came out, it was gone.

Infer that Jack will be angry.

In Section 2, we describe the logic used in this paper. In Section 3, we outline a proof of the desired inference. In Section 4 we contrast our theory with previous work, and in Section 5 we present our conclusions and discuss future work.

2 The logic

2.1 Syntax

In this paper, we use an extension of the temporal logic developed in [Shoham 1988], modified to incorporate the three modal operators 'want,' 'know,' and 'believe.' We use an S5 axiom set for 'know,' weak S5 for 'believe,' T without veridicality for 'want,' and the inference rules modus ponens and universal instantiation (cf. [Hughes & Cresswell 1968]). Unlike Shoham, we assume that the intervals over which an assertion is interpreted are closed. Shoham deliberately makes no commitment one way or the other; we have found that, for purposes of our proofs, it is useful to make a choice, and in general, closed intervals seem more intuitive.

The syntax of our language is the same as in [Shoham 1988], with the following additions:

1. The symbols in the language include the three modal operators, 'want,' 'know,' and 'believe.'
2. If trm_a and trm_b are temporal terms and ϕ is a modal formula, then TRUE(trm_a, trm_b, ϕ) is a formula.

   The set of *modal formulas* is defined as follows:

   (a) If O is one of the three modal operators, x is a nontemporal term denoting some person, r is an n-ary relation symbol, and $trm_1, ..., trm_n$ are nontemporal terms, then $(O\ x\ (r\ trm_1 ... trm_n))$ is a modal formula.

   (b) If O is one of the three modal operators, x is a nontemporal term denoting some person, and ϕ is a modal formula, then $(O\ x\ \phi)$ is a modal formula.

2.2 Semantics

2.2.1 Informal semantics

Intuitively, the semantics of our language can be understood as follows. We are given a set that includes all of the possible worlds. Possible worlds have a temporal dimension. That is, each possible world is a complete possible history of the world, like a timeline extending infinitely far into the past and the future.

In all the worlds that are knowledge-accessible to an agent x, the propositions that x knows about the past, present, and future all hold. All other propositions vary from world to world. Similarly, in all the worlds that are belief-accessible to x (which may not include the 'real' world, if x has beliefs that are inconsistent with reality) all of x's beliefs hold. Propositions about which x has no particular opinion vary from world to world. Finally, the propositions that x wants to be true are true in all the 'want-accessible' worlds, the propositions x wants to be false are false, and propositions about which x is indifferent vary from world to world.

2.2.2 Formal semantics

Formally, the semantics of our language are as follows:

Let D be a domain of individuals, and let $P \subset D$ be a (nonempty) subset of D consisting of all of the persons in D. Let PW be the set of all possible worlds. With each $x \in P$ we associate three relations on PW, B_x, W_x, and K_x, corresponding to the modal operators 'believe', 'want', and 'know', respectively. Let O represent any one of the three modal operators, and let O_x represent any of the three relations. Each of these relations is serial, i.e., it has the property that from any given world at any time, at least one other world is accessible: $\forall w_i \in PW, \forall t_i(\exists w_j \in PW(O_x\ w_i\ w_j\ t_i))$. An interpretation I is a function that maps the nonlogical symbols in the language to some element of D.

Given these definitions, a sentence ϕ is true in a world $w_i \in PW$ under an interpretation I and a variable assignment VA if and only if one of the following is true:

1. ϕ has the form $trm_1 = trm_2$ and $I(w_i, trm_1) = I(w_i, trm_2)$.
2. ϕ has the form $trm_1 \preceq trm_2$ and $I(w_i, trm_1) \leq I(w_i, trm_2)$.
3. ϕ has the form $\mathrm{TRUE}(trm_a, trm_b, (r\ trm_1 ... trm_n))$ and the relation $I(w_i, I(trm_a), I(trm_b), r)$ holds on $I(w_i, I(trm_a), I(trm_b), trm_1)$ through $I(w_i, I(trm_a), I(trm_b), trm_n)$.
4. ϕ has the form $\mathrm{TRUE}(trm_a, trm_b, \psi)$, where ψ is a modal formula, in the form $\psi = (O\ x\ \zeta)$, O_x is the relation corresponding to O, and $\mathrm{TRUE}(trm_a, trm_b, \zeta)$ holds in all worlds w_j such that $\forall t_i, trm_a \leq t_i \leq trm_b, (O_x\ w_i\ w_j\ t_i))$.
5. ϕ has the form $\phi_1 \wedge \phi_2$, and both ϕ_1 and ϕ_2 are true.
6. ϕ has the form $\neg\phi_1$, and ϕ_1 is false.
7. ϕ has the form $\forall z \phi_1$, and ϕ is true under all variable assignments VA′ that agree with VA everywhere except possibly on z.

2.2.3 Basic concepts

As stated above, we use some key concepts from deontic logic: 'permitted', 'prohibited', and 'obligated' [Wright 1951]. 'Permitted' is a predicate on actions. Thus, the proposition '(permitted rules (do x a))' is true in a given world if and only if the relation corresponding to the symbol 'permitted' holds on the elements corresponding to 'rules' and '(do x a)' in that world, that is, if x is in fact permitted by the body of rules in question (e.g., law or ethics) to perform that action. For example, we might have ¬(permitted Law (do Antigone (bury Polynices))) and ¬(permitted Religion (do Antigone ¬(bury Polynices))). Permission might also be granted by an individual, for example, (permitted Wilma (do Jack (take car))).

'Obligated' and 'prohibited' are defined in terms of 'permitted' in the usual way. An action is prohibited if performing it is not permitted. An action is obligated if *not* performing it is not permitted. Where the source of the rules is religion or ethics, these predicates might be expressed in terms such as 'right' and 'wrong', or 'should' and 'should not.' The more general formulation allows us to express a variety of types of rules and prohibitions using a single set of predicates.

anger $\forall x, t_1 \leq t_2, t_3 \leq t_4, y, a,$
[TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (occurs (do y a))))) $\wedge$
TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (obligated Ethics $\neg$(do y a))))) $\wedge$
TRUE(t_1, t_2, (want x TRUE($t_3, t_4, \neg$ (occurs (do y a))))) $\wedge$
TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (know y TRUE(t_3, t_4, (obligated Ethics $\neg$(do y a)))))))
$\mapsto$
TRUE(t_1, t_2, (angry x y (do y a) t_3 t_4))]

You become angry at someone if you think they did something wrong, you didn't want them to do it, and you think they knew it was wrong.

gratitude $\forall x, a, t_3 \leq t_1 \leq t_2, t_3 \leq t_4 \leq t_2, y \neq x,$
[TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (occurs (do y a))))) $\wedge$
TRUE(t_1, t_2, (want x TRUE(t_3, t_4, (occurs (do y a))))) $\wedge$
TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (benefit x (do y a))))) $\wedge$
TRUE(t_1, t_2, (believe x $\neg$ TRUE(t_3, t_4, (conditional (do y a)))))
$\mapsto$
TRUE(t_1, t_2, (grateful x y (do y a) t_3 t_4))]

You are grateful to someone if you think they did something that you wanted them to do that benefited you, and their action was not conditioned on receiving anything in return.

approval $\forall x, y, t_1 \leq t_2, t_3 \leq t_4, a,$
[TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (occurs (do y a))))) $\wedge$
TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (obligated Ethics (do y a)))))
$\mapsto$
TRUE(t_1, t_2, (approve x y (do y a) t_3 t_4))]

You approve of someone if you believe that they have done something they should.

disapproval $\forall x, y, t_1 \leq t_2, t_3 \leq t_4, a,$
[TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (occurs (do y a))))) $\wedge$
TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (obligated Ethics $\neg$(do y a)))))
$\mapsto$
TRUE(t_1, t_2, (disapprove x y (do y a) t_3 t_4))]

You disapprove of someone if you believe that they have done something they shouldn't.

shame $\forall x, t_1 \leq t_2, t_3 \leq t_4 \leq t_2, a,$
[TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (occurs (do x a))))) $\wedge$
$\exists y \neq x,$
[TRUE(t_1, t_2, (believe x TRUE(t_1, t_2, (disapprove y x (do x a) t_3 t_4)))) $\wedge$
TRUE(t_1, t_2, (want x TRUE($t_1, t_2, \neg$(disapprove y x (do x a) t_3 t_4))))]
$\mapsto$
TRUE(t_1, t_2, (ashamed x (do x a) t_3 t_4))]

You feel ashamed if you believe that you have done something that someone else thinks is wrong, you think they know what you've done, and you care what they think.

guilt $\forall x, t_1 \leq t_2, t_3 \leq t_4 \leq t_2, a,$
[TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (occurs (do x a))))) $\wedge$
TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (obligated Ethics $\neg$(do x a))))) $\wedge$
TRUE(t_1, t_2, (believe x TRUE(t_3, t_4, (know x TRUE(t_3, t_4, (obligated Ethics $\neg$(do x a)))))))
$\mapsto$
TRUE(t_1, t_2, (guilty x (do x a) t_3 t_4))]

You feel guilty if you believe that you have done something that you think is wrong, and you believe that you knew it was wrong at the time you did it.

Figure 1: Definitions.

Therefore, immediately after Jack parked the car, it was at the parking place. At a certain later time, the car was not at the parking place. Things don't move by themselves. Therefore, someone must have moved it to another place. When things are moved, they are then at the location to which they were moved. Therefore, the car was then at the place to which it was moved.

Jack was at the parking place at a later time when his car was not there. People can't be in two places at once. Therefore, Jack was not at the place where the car was.

Jack owned the car. Jack did not possess the new place where the car was. If you move something out of someone's possession, you have taken it away from them. Therefore, someone took Jack's car between the time he parked it and the later time when it was not in its parking place.

Jack had not given anyone permission to take his car. Therefore, Jack had not given the person who took his car permission to take it. The law permitted Jack to park the car where he did (i.e., it was a legal parking place.) The law doesn't permit anyone to take a car unless it's parked in an illegal location, and then only if that person is an official. Therefore, the law did not permit whoever took Jack's car to do so.

The person who took Jack's car did not own the car. It is wrong (ethically prohibited) to take something that you do not own without the owner's permission. Therefore, it was wrong for the person who took Jack's car to do so.

Jack believes the axioms and can reason. Therefore, Jack believes that someone took his car, and that it was wrong for them to do so. If you own something and you haven't given anyone permission to take it, you don't want anyone to take it. Therefore, Jack didn't want anyone to take his car. Therefore, he was not grateful to the person who took his car.

All agents believe the axioms and can reason. Therefore, whoever took Jack's car knew it was wrong. Jack believes that whoever took his car knew it was wrong. It follows, therefore, that Jack is angry at the person who took his car.

4 Previous Work

Little previous work has been done in this area. Kube encodes a portion of commonsense psychology in [Kube 1985]. That paper is restricted to the ways in which people acquire knowledge and beliefs, however, and explicitly excludes any consideration of agents' emotions or intent.

Dyer attempted to incorporate psychological knowledge in his program BORIS [Dyer 1983]. BORIS's approach is substantially less formal than ours. It processes one basic story using some fairly simple definitions for the emotions involved. According to these definitions, for example, you are happy if you achieve a goal; if one of your goals fails or is suspended, you will be unhappy; and if someone else causes this to happen, you will become angry. In general, these goal-oriented definitions are too broad. For example, if you go to the bank and there is a long line at the teller machine, the people in front of you are causing one of your goals to fail – the goal of obtaining cash quickly. BORIS's definition would sanction the inference that you are angry at all of those people. Our definition would allow this inference only if you believe that their actions are wrong. You might become angry at someone who cut in front of you in line, but not at someone who merely arrived a few minutes before you.

Lehnert uses the concept of 'affect states' in her work on plot units, but makes no attempt to describe complex emotional states in detail. For her purposes, it was only necessary to distinguish between positive events, negative events, and neutral mental states [Lehnert 1982].

There is a substantial psychological literature on the emotions (See, e.g., [Strongman 1987] and works cited therein.) In general, however, this research disregards the kind of commonsense knowledge we are trying to encode. Instead, it focuses on such issues as the possible neurological causes of emotion. Rather than using a commonsense theory as a basis, even when addressing some of the same issues, this work usually takes an independent approach [Harre et al. 1985].

For our purposes, the most interesting of the psychological theories of emotion is the recent work by Ortony, Clore, and Collins [Ortony et al. 1988]. They attempt to characterize the range of possible emotions, rather than the emotional reactions which would be likely to occur in a given culture. They provide a broad framework which is generally consistent with our theory. Like our theory, theirs assumes that emotions are largely caused by people's beliefs about the world. They divide these beliefs into three categories – beliefs about actions, events, and objects – and divide emotions into three basic types, according to the kind of beliefs by which they are triggered. Thus, for example, joy is a positive response to an event, approval is a positive response to an action, and liking is a positive response to an object. All of the emotions considered in this paper would be classified as responses to actions; however, our theory could be extended to handle the other areas as well.

Unlike our theory, Ortony et al.'s work is very informal. Their terminology has no precise semantics. The reader must rely on intuition for the meaning of terms such as 'joy-intensity' and 'fear-potential.' Our theory introduces very few new primitives; most predicates are defined in terms of a small group of well-studied logical concepts. Cain and O'Rorke have begun working on a story-comprehension system based on Ortony et al.'s theory [Cain & O'Rorke 1988]. Because of the breadth of the underlying theory, their system promises to be more general than BORIS; however, like BORIS, this work should be considerably less formal than ours.

5 Conclusions and Future Work

In this paper, we define a logic in which commonsense knowledge about human psychology can be expressed. We demonstrate, by formulating and solving a set of benchmark problems, that our logic is sufficiently expressive to handle a variety of information about human actions and responses, in a way that is substantially more formal than previous work in the area.

Obvious extensions to this work include defining further emotions, such as hope, fear, surprise, and impatience, along with their causes and results. Because our theory is compositional, some extensions fall easily out of the definitions. For example, we could define an emotion that might be labelled 'remorse' – the response when you realize that you have done something wrong, although you didn't know it was wrong at the time – by modifying the definition of guilt only slightly. Other extensions could be obtained by incorporating additional predicates in the theory, perhaps following the directions suggested by [Ortony et al. 1988].

In addition, we would like to integrate this work with a theory of causation. Note that our rules purport to define the 'causes' of states such as anger, gratitude, shame, and guilt. For the purposes of this paper, we have treated the causal inferences as though they were equivalent to logical inference. Technically, however, causation is both stronger and weaker than logical inference (see discussion in [Shoham 1988], pp. 166ff). Ideally, a theory of emotion should incorporate a more precise understanding of causation.

Finally, we would like to explore an issue that is implicit in the definitions of approval and disapproval. Note that these definitions imply a symmetry between approval and disapproval that

does not in fact exist. You might disapprove of someone for committing murder (say), but you do not constantly approve of everyone who is refraining from murder. Similarly, you might approve of someone who makes a large donation to charity, while not disapproving of those who do not.

This asymmetry results from a fact which holds for the other emotions as well: typically, we only react to things which we do not take for granted. You do not approve of everyone who fails to commit murder, unless the temptation is particularly severe, because you take that for granted. Similarly, children are not always grateful for the food, clothing, and shelter provided by their parents, because they take these things for granted. We only react to actions that we notice, for one reason or another. Possibly this could be tied in with notions of awareness and implicit/explicit knowledge from epistemic logic [Fagin & Halpern 1985]. Developing a general theory of what kinds of things we take for granted is outside the scope of this paper, but remains an interesting area for future work.

6 Acknowledgements

This research has been supported by NSF grants IRI 8515005 and IRI 8801253. The author gratefully acknowledges the assistance of Leora Morgenstern, Tom Dean, Eugene Charniak, Robert McCartney, and Mary Harper.

References

[Cain & O'Rorke 1988] Cain, Timothy, and Paul O'Rorke. Explanations involving emotions. *Proc. AAAI-88 Workshop on Plan Recognition.*

[Davis 1988] Davis, Ernest. Inferring Ignorance from the Locality of Visual Perception. *AAAI-88.*

[Dyer 1983] Dyer, Michael G. *In-Depth Understanding.* Cambridge, MA: MIT Press (1983).

[Fagin & Halpern 1985] Fagin, Ronald and Joseph Y. Halpern. Belief, Awareness, and Limited Reasoning: Preliminary Report. *IJCAI-85.*

[Harre et al. 1985] Harre, Rom, et al. *Motives and Mechanisms: an introduction to the psychology of action.* London: Methuen (1985).

[Hayes 1985] Hayes, Patrick. Naive Physics I: ontology for liquids. In *Formal Theories of the Commonsense World,* J. Hobbs and R. Moore, edd., Ablex 1985.

[Hughes & Cresswell 1968] Hughes, G.E. and M.J. Cresswell. *An Introduction to Modal Logic.* London: Methuen (1968).

[Kube 1985] Kube, Paul. Cognitive propositional attitudes. In *Commonsense Summer: Final Report.* Stanford University: Center for the Study of Language and Information, Report No. CSLI-85-35 (October 1985).

[Lehnert 1982] Lehnert, Wendy. Plot units: a narrative summarization strategy. In *Strategies for Natural Language Processing,* W. Lehnert and M. Ringle, edd., Hillsdale, NJ: L.E.A. (1982).

[McDermott 1982] McDermott, Drew. A Temporal Logic for Reasoning About Processes and Plans. *Cognitive Science 6:* 101-155 (1982).

[Ortony et al. 1988] Ortony, Andrew, et al. *The Cognitive Structure of Emotions.* Cambridge: Cambridge University Press (1988).

[Sanders 1989] Sanders, Kathryn E. *A Logic for Emotions: a basis for reasoning about commonsense psychological knowledge.* Tech. Rep. 89-23, Brown University Computer Science Dept.

[Shoham 1988] Shoham, Yoav. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* Cambridge, MA: MIT Press (1988).

[Strongman 1987] Strongman, K. T. *The Psychology of Emotion.* Wiley (3d. ed. 1987).

[Wright 1951] Wright, G. H. von. *An Essay in Modal Logic.* North-Holland (1951).