

Human XP: Representing Belief, Desire, and Hidden Meanings

Jerry Weltman

Louisiana State University
Baton Rouge, Louisiana 70802 USA
jweltem2@lsu.edu

ABSTRACT

The recently proposed Human Experience Project seeks to collect a large corpus of scenes about everyday life. The project calls for volunteers to create 3-D scenes in a Virtual World environment and annotate them with information that can be used in commonsense modeling. The annotations label the objects in a scene, the various actions that occur, and *why* they occur. To help explain the why behind everyday activities, the project must represent complex mental states. Presented here is the project's approach to representing belief, desire, and the hidden meanings of speech acts.

Author Keywords

knowledge acquisition, common sense, mental states.

ACM Classification Keywords

I.2.6. Learning: Knowledge Acquisition.

INTRODUCTION

Virtual Worlds can teach a computer a lot about common sense. Environments such as Second Life [6] contain detailed models of buildings and landscapes of a world similar in many ways to our own. Animation software tools such as Blender [2] produce scenes with life-like actors in realistic 3-D settings. These virtual world environments implicitly provide valuable commonsense information about how things move, what they look like, and how they interact. However, there is currently no framework for making scenes that are more useful for story understanding. To build such a framework, I propose the Human Experience Project, or Human XP.

The goal of Human XP is to collect a large corpus of annotated 3-D scenes of everyday life. This corpus would provide researchers with valuable raw data for designing neural networks, semantic networks, statistically-based rules, or other structures for artificial intelligence and natural language processing such as ThoughtTreasure[5]. Similar to collaborative commonsense projects at the MIT Media Lab [4,7], in particular ComicKit [8], Human XP will rely on volunteers to contribute simple narratives using a web-based interface. However, Human XP is based on two novel components: 1) animated scenes in a Virtual World environment; and 2) detailed annotations that describe objects, actions, mental states, and hidden

meanings behind speech acts. A large challenge of Human XP is to provide an integrated framework that will make it easy for amateurs to describe objects, actors, what they are doing, and why they are doing it.

I have previously introduced this project and presented the broad requirements [10]. Here, I will focus on the part of the Human XP framework that represents beliefs, feelings, and hidden meanings. But first let's see an excerpt of a simple annotated scene.

EXCERPT FROM A SCENE

The scene is called "Max breaks a vase." It begins with Max feeling bored, having nothing to do, sitting in a living room. He notices a lovely vase, walks over to it, picks it up, and drops it, smashing it into pieces. It is great fun! The scene shows Max's mental state before breaking the vase and what happens as a result of his action.

A scene consists of one or more frames. Each frame has two areas. On the left side is a 3-D depiction of a time slice of Max's life. On the right side is the "maxometer," a measure of Max's perceptions and mental states that directly result from the most recent action. The maxometer categorizes feelings according to whether they are positive, negative, or neutral. The relative intensity of each feeling is represented by a corresponding slider bar. All of the frame data, including the actions on the left and the maxometer data on the right, are specified by a Human XP volunteer contributor. Figure 1 depicts Max noticing the vase on a table. This action causes Max to have a curious feeling, as indicated by the maxometer. He gets up and walks to the vase. Figure 2 shows him grasping the vase to pick it up.

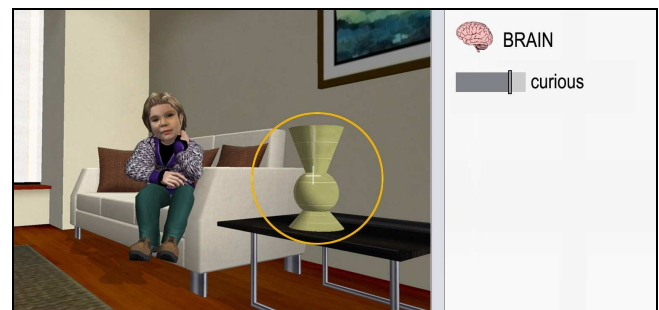


Figure 1: Max sees the vase

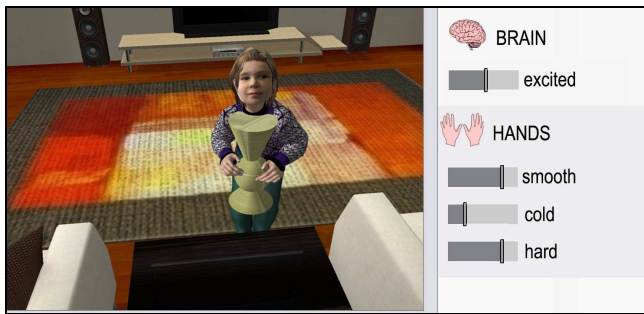


Figure 2: Max grasps the vase

The maxometer shows that his hands feel the pleasant sensation of the ceramic that he is about to smash. His mental state has moved from curiosity to excitement.

Creating a scene will require a software tool which I call Scene Builder. In many respects, Scene Builder will be similar to current 3-D animation and modeling software that allow users to put objects into a scene to create an animated movie. However, unlike a passively viewed movie, Human XP scenes are designed to be viewed in a Virtual World environment, which allows the audience to manipulate the scene to view it from any angle. Also, the annotations (which consist of scene labels, object labels, and maxometer readings) are as important to the creative process as the scene itself.

REPRESENTING MENTAL STATES

A long-held goal of Artificial Intelligence is to create programs that understand stories, even simple children's stories. Consider the following story fragment:

Mommy and Max are playing hide-and-peek. Mommy has hidden behind the couch. Max notices a bump in the curtain. He tiptoes up to the curtain and grabs it. Excitedly, he pulls back the curtain but is surprised to see nothing there.

Why is Max excited when he pulls back the curtain? Why is he surprised to see nothing there? To build a story-understanding program capable of answering these "Why" questions, it seems critical that AI researchers construct models of human behavior that incorporate propositional attitudes – intentional mental states such as beliefs and directed desires. Annotations that explain why an actor is excited or surprised would contribute towards this research, but there are significant challenges in how to structure these annotations. How do we represent the full scope of what a person believes at a particular moment so that we can justify a surprise? How do we represent the belief states of different actors in a scene when each actor has his/her own point of view? How do we represent expectations so that we can explain excitement? The next subsections address these issues.

Representing Belief

In Human XP, the scene itself represents one actor's beliefs about the setting and actions of a scene. For example, in the

"Max breaks the vase" scene, Max's internal belief state is represented by the living room and all the objects in it. Obviously, Max has beliefs unrelated to the living room, but here we seek to represent only the beliefs most relevant to the current scene. The living room has a couch, table, vase, picture on the wall, etc, and this means that Max *believes* that the living room contains these objects. If Max is unaware of the picture on the wall, then it should not be depicted in the scene, even if the scene creator knows that the picture exists. As with objects, any action that occurs in the scene is part of Max's belief state. If something occurs in the room that Max does not notice (e.g. someone sneaks up behind Max), the action would not be represented in the scene. In sum, a belief state of one actor is represented directly by the actions and the objects in a scene. But what happens if there are conflicting belief states between actors?

In the previous hide-and-peek story, Max believes Mommy is behind the curtain even though she is really somewhere else (in fact, she is hiding behind the couch). An accurate representation of Max's belief would show Mommy behind the curtain. On the other hand, a representation of Mommy's belief would show her behind the couch. Thus, we have conflicting belief states. To model different belief states, Human XP allows the same time-slice of events to be depicted in multiple views. There could be a "Max View" of the scene and a "Mommy View." There can also be an objective "Audience View" which would show what is happening as if an audience were witnessing the scene.

When the same events have multiple viewpoints, they are modeled as separate parallel scenes. Figure 3 shows a fragment of a hide-and-peek scene. On the left side is the Audience View, with two frames. Each frame is labeled with a unique ID number that is chosen by the Human XP framework. In frame 1000.1 Max grabs the curtain; in 1000.2 Max opens the curtain. In both frames, Mommy is shown hiding behind the couch. The Audience View is omnipresent, which means the scene can be viewed from all angles. The Virtual World capability of modeling all angles is important when representing belief states. It allows Human XP contributors to represent objects that are not visible by the audience or actors but still exist in their internal representation of the world.

On the right side of figure 3 is the Max View, which models Max's belief state. The first frame of this scene has ID 1200.1. Mommy's figure stands where Max believes she is actually located – behind the curtain. The maxometer shows that Max expects to find Mommy, he is excited and having fun. In general, the mental states are assumed to form a causal chain. Max's expectation causes him to be excited; his excitement causes him to have fun.

In frame 1200.2, the maxometer shows Max's surprise. The figure of Mommy is shown with a 0% next to it. By default, an actor believes 100% of everything in the scene. A label of 0% next to Mommy shows that Max is actively thinking about the fact that Mommy is NOT at that location. The

same percentage notation can represent beliefs such as *maybe* (< 50%) and *probably* (> 50%).

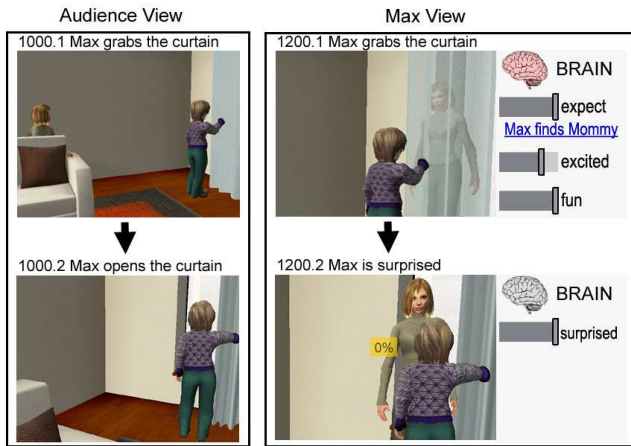


Figure 3: Conflicting belief states

We have seen how Human XP represents belief states of a single actor, and how conflicting belief states can be represented by multiple views. The next subsection addresses representing desires and expectations.

Representing Desire and Other Counterfactuals

Mental states such as desire, hope, expectation, and fear refer to counterfactual events – events that have not happened and may never happen. As we have seen, the maxometer in frame 1200.1 shows an EXPECT “[Max finds Mommy](#)” mental state. To capture commonsense information about what Max expects, the scene “Max finds Mommy” must be represented even though it will never happen. (Remember, Mommy is behind the couch.) Otherwise, there will be no data to explain why Max is excited when he grabs the curtain.

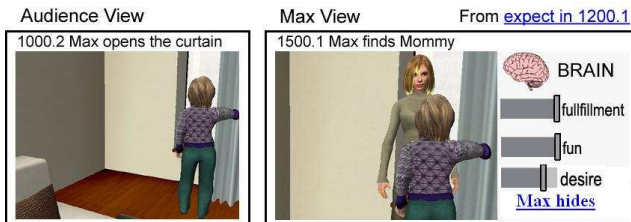


Figure 4: Represented an expected event

Figure 4 shows the “Max finds Mommy” scene, parallel with what actually happens when Max opens the curtain. In contrast with the Max View of Figure 3, this version of Max View has the annotation “From [expect in 1200.1](#)” This annotation identifies the scene as a counterfactual, invoked from an expectation rather than a depiction of a real event. Reality has not changed in the left side of the figure; Mommy is still behind the couch. The counterfactual scene simply shows what Max *expected* to happen at that time. He expected to feel a strong sense of fulfillment and fun from having found Mommy. He also expected to desire to play another game in which he hides and Mommy searches.

The DESIRE “[Max hides](#)” in Figure 4 is an example of a counterfactual that links to another counterfactual. As Max was expecting to find Mommy, he was already anticipating wanting to play a new game where he hides. Thus, his expectation has a desire within it. This desired scene is not shown here, but it would simply be a counterfactual scene, invoked from an expected scene.

Scene references behave like hyperlinks in a web browser; they allow users to jump from a mental state in the maxometer to a scene associated with that state. Each scene, whether it is believed to be true or it is an unrealized desire, is anchored to a specific time, as shown in Figure 5.

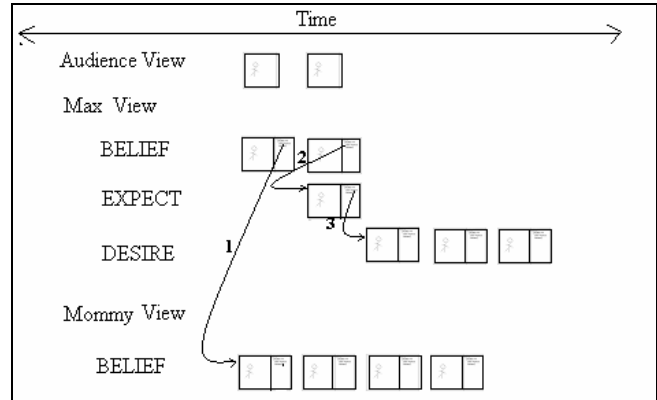


Figure 5: Global view of related scenes

The figure depicts the hide-and-seek story discussed above, with arrows representing the links from a mental state to its associated scene. The two frames of the Audience View are on top. Next, are the frames of the three Max Views: the first is what Max believes to be happening. The second (linked by arrow 2) is a counterfactual: it is what Max expects to happen when he pulls back the curtains. The third (linked by arrow 3) is also counterfactual: it represents Max’s desire to play another game after finding Mommy.

Arrow 1 of Figure 5 depicts another possible reference, not mentioned yet. Suppose Max believes that Mommy is having fun as well. To show Mommy having fun, we show Mommy’s maxometer with a FUN mental state. Thus, we must show the scene from her point of view. There can be multiple types of Mommy Views: one view could be what Mommy herself believes; another view could be what Max *thinks* Mommy believes. Because it is invoked from Max’s belief, the Mommy View in Figure 5 represents this second type of Mommy View. This type of view is an example of first order Theory of Mind, the ability to attribute feelings and perceptions to others. Theory of Mind views are important to explain human behavior, and are particularly useful in annotating scenes with dialog because they help show what the speaker intended and what the hearer understood.

SCENES WITH DIALOG

Human XP uses dialog bubbles to represent speech. However, simple bubbles cannot capture many prosodic

and paralinguistic features which are critical to understanding speech. Therefore, a contributor will be able to attach annotations such as tone of voice (e.g. ANGRY, JOKING), volume (e.g. WHISPER, LOUD), and speech rate (e.g. FAST, SLOW). Traditional mark-up such as punctuation and underscored words will also be used.

One important goal in representing dialog is to provide data for the study of Pragmatics. Even a simple dialog is full of hidden meanings springing from Pragmatic concepts of presuppositions, implicatures, and illocutionary acts. For example, “Max is happy that it is raining” presupposes that it is raining; “Can you pass the salt” carries the general implicature of a request; and “What a beautiful day!” has the illocutionary force of an assertion that the weather is beautiful. Human XP allows participants to express hidden meanings behind spoken utterances, as well as how each actor interprets an utterance.

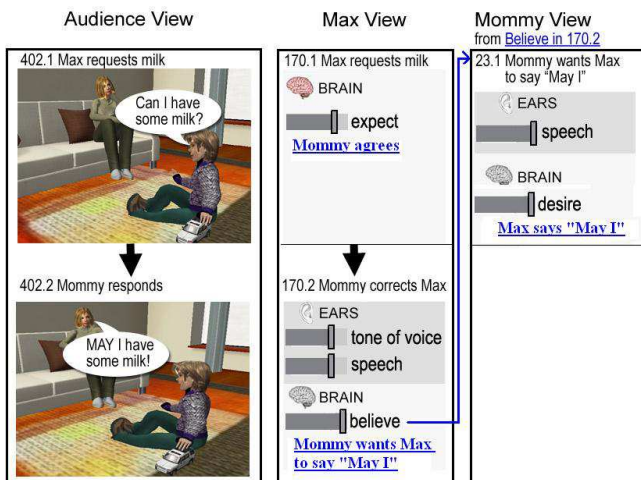


Figure 6: Hidden meanings behind dialog

The mechanism for exposing hidden meanings behind dialog utterances is the maxometer. For example, before Max says to Mommy “Can I have some milk”, the maxometer would show Max is thirsty and desires that Mommy bring him milk. Human XP contributors will decide the meanings of the utterances and annotate them accordingly.

Figure 6 shows a fragment of a scene with dialog. Three views are represented: Audience, Max, and Mommy. The actions of the scene are shown only once, in the Audience View. For convenience, the other two views are shown with just a maxometer. In the Audience View, Max requests milk by saying “Can I have some milk?” Mommy’s response is to correct Max’s grammar.

The maxometer in the Max View shows that Max expects Mommy to agree to his request. The reference “[Mommy agrees](#)” is to a counterfactual scene not shown in the figure. When Mommy responds, Max attends to her utterance. The combination of the stern tone of voice and the words of the speech cause Max to believe that Mommy wants him to say

“MAY I have some milk”. Max’s belief about what Mommy desires is represented by a link to a scene from Mommy’s point of view. Since it is linked from Max’s belief, it represents what Max *believes* about Mommy.

CONCLUSION

Representing and reasoning about mental states is an important aspect of AI [1],[3],[9]. ComicKit uses the familiar thought bubbles of comic strips to represent mental states, but the Human XP proposal is far more ambitious. By using scenes in a Virtual Worlds and a new visual framework, Human XP seeks to capture a fuller range of the human experience, including perceptions, feelings, beliefs, desires, and the Pragmatic meaning behind utterances.

ACKNOWLEDGMENTS

I am indebted to Margit Link-Rodrigue, who created the scene graphics and provided many valuable suggestions.

REFERENCES

1. Bacon, W. F. What everyone knows about attention. *Representing Mental States & Mechanisms: Papers from the 1995 AAAI Symposium*. Edited by Michael T. Cox & Michael Freed, co-chairs. Technical Report SS-95-05.
2. Blender. Open Source 3D Graphics Creation. <http://www.blender.org/>
3. Dyer, M.G. *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*, MIT Press, Cambridge, MA, 1983.
4. Leiberman, H., Smith D., and Teeters, A. Common Consensus: A Web-based game for collecting commonsense goals. Presented at the Commonsense Workshop at IUI-07, 2007.
5. Mueller, E.T. *Natural language processing with ThoughtTreasure*. New York: Signiform, 1998.
6. Rymaszewski M., Au W.J., Wallace M., Winters C., Ondrejka C., Batstone- Cunningham B. *Second Life: The Official Guide*, Wiley, 2006.
7. Speer, R. Open Mind Commons: An inquisitive approach to learning common sense. In *Proc. Workshop on Common Sense and Interactive Applications*, 2007.
8. Williams, R., Barry, B., Singh, P. ComicKit: Acquiring story scripts using common sense feedback. In *Proc. IUI-05*, 2005
9. Wooldridge, M. *Reasoning About Rational Agents*, MIT Press, Cambridge, MA, 2000
10. Weltman, J. Human XP: Using virtual worlds to capture common sense. *Proc. of Mardi Gras Conference on Virtual Worlds*, 2009 (forthcoming)