Empirical Knowledge Acquisition of Commonsense Psychology

William Jarrold & Peter Z Yeh
Nuance Communications
william.jarrold@nuance.com peter.yeh@nuance.com

Abstract

A method for acquisition of general inference rules in the social / emotional domain is presented. This method is characterized as an empirical, psychologically-guided means of developing and evaluating models of this type of reasoning. The focus is the task of predicting the range of emotions and *explanations* for each emotion felt by agents in a range of given scenarios. Results obtained from implementing this method for a set of simple and complex scenarios are described. The products of this pilot study are acquired knowledge in the form of a simple rule-based model plus a small corpus of training and evaluation data. Extensions to the method are proposed to support incremental knowledge acquisition which form a generate-and-test loop between modelers and crowdsourced laypeople.

1 Introduction

Commonsense psychology is a critically necessary capability that facilitates natural and fruitful interaction with humans. The term *affective theory of mind* (ATOM) refers to the specific capacity to predict, explain and understand others emotional appraisals. It is considered a sub-capacity of theory of mind (TOM) - the ability to reason about the mental states of others [Ravenscroft, 2010]. TOM has overlapping meaning with terms such as "Social-emotional intelligence", "folk psychology" and "mindreading". Without this capacity effective social interaction and communication would be impossible.

An important principle motivating our ATOM-focused approach is that the *explanation* for an inference or prediction is important. A system which can explain its predictions is more understandable. As the workshop announcement states "Knowledge should be representable in a form understandable by humans".

Another motivation is derived from a recent resurgence of interest in evaluation of human-level AI. A main theme of this work has been in remediation of methodological challenges with the Turing test [Levesque, 2011]. A main theme in this resurgence is that instead of a single evaluation in the classic Turing Test there should be a several different types of tests each with incremental gradations in difficulty focused on particular domains or dimensions such as vision, robotics, etc. [Adams *et al.*, 2016]. In this vein the Social-Emotional Turing Challenge [Jarrold and Yeh, 2016] proposes a five-stage method involving incremental evaluations in affective reasoning that are difficult to game. Many aspects of the method in the aforementioned publication are similar to the method described below. However, in the previous paper the focus was *assessing* human-level ATOM whereas in the present paper we present additional results and modify the method for the *acquisition* ATOM knowledge.

An additional motivation derives from the need to go beyond introspection and develop methods to involve empirical validation and evaluation in the knowledge acquisition process. Commonsense reasoning for other domains tends to have been pursued using introspection. Consider, for example, the introspective nature of the seminal work in commonsense reasoning about naive physics (e.g. [Hayes, 1985]). Although this may work for precise and objective domains such as naive physics it is less appropriate for less predictable or hard to "pin down" domains like naive psychology.

Perhaps some of the hard to pin down nature of commonsense psychology stems from the fact that there are many viable interpretations to a given situation. This paper starts from the assumption that although there a variety of reasonable yet incompatible predictions one might make, there is still some systemiticity. This systemiticity is embodied in the *Generativity Hypothesis* i.e. that there are a multitude of appraisals that can be generated for a given scenario but that the predictions of ATOM are falsifiable.

Such a model or theory needs to be able to reproduce the range of appraisals that a group of typical humans collectively generate when thinking about how other humans might feel in a given situation. It should exhibit enough generativity so as to approximate the types of variability that can be seen in human cognition about others' emotional reactions [Ortony, 2001]. Yet the theory should also be restrictive in that it should predict that certain types of appraisal are inappropriate in certain situations. In sum, what we are after is a theory that produces diverse yet falsifiable predictions. To put it another way, such a model should be analogous to a generative grammar - producing all and only viable appraisals of a given situation.

In addition to being generative, the theory must be shared

between most people, otherwise how would we understand each other? How would we be able to make reasonably accurate predictions about how another was feeling?

Based on these motivational tenets a method is described below that can be applied towards the acquisition of ATOM knowledge for agents.

2 Background

A variety of prior work has addressed the problem of acquisition of knowledge relevant to commonsense psychology. [Ortony *et al.*, 1988] is well-known starting point for the computational modeling of emotion because computational tractability was one of the goals of the authors. The theory describes a broad ontology of emotion appraisal, articulating how an individual's desires, standards, and values play into the appraisal a situation in terms of twenty two distinct emotion types. However, there are some gaps. First although the authors make ontological claims about a set of concepts and relations for describing appraisal structure, there is little in the way of axiomatization. Although the theory may provide an overall appraisal architecture, it does not provide specific guidance on how representations of a given agent's specific desires, standards and values combine with real world events and objects so as to produce an appraisal resulting in the predicted experience of an emotion. As a result, more work remains to be done that will enable a system to actually generate appraisals given a particular character's situation.

Hobbs, J. and Gordon, A. (2011) builds upon the work of [Ortony *et al.*, 1988]. It addresses some of the above gaps because the authors state that they have axiomatized the OCC theory presumably providing sufficient conditions for each of the emotion types. However, there is only one such example axiom provided in the paper. In addition, it appears to be derived via introspection. There is no implementation or evaluation and thus we do not know how believable its predictions and explanations are.

[Gordon, 2016] addresses many of these limitations. The paper describes the evaluation of a model on a benchmark task requiring a system to interpret observable psychological states (including emotion and intention) given a description of agents in a situation. This benchmark task, called Triangle-COPA [Maslan *et al.*, 2015], is based on early work in human social perception [Heider and Simmel, 1944] in which subjects were presented with a short animated film depicting the movements of two triangles and circle in and around a box with a hinged opening. Despite the simplicity of the film, subjects produced surprisingly rich narratives which anthropomorphized the moving shapes as intentional characters with beliefs, goals, emotions, and social relationships.

Taking inspiration from the [Heider and Simmel, 1944] film and its associated experimental task, the Triangle-COPA benchmark consists of one hundred items. For each item, there is a description of short sequence of events involving the above mentioned geometric figures rendered in English and first order logic. This description is followed by a question requiring an interpretation of the action sequence. Finally, there are two possible interpretations provided. The task is to choose the "correct" interpretation which been determined unanimously by multiple human raters.

[Gordon, 2016] describes a hand-coded first-order logic model that solves 91 of the 100 questions in the Triangle-COPA benchmark. This meets many of the desiderata outlined above because it is axiomatized, makes predictions about an agent's psychological states and is subject to evaluation. However, it falls short of the problem circumscribed in the introduction for a couple of reasons. First, the domain of triangle and circle motions is, of course, a much more narrow one than the domain of human characters acting in realistic worlds. A second related point is that the models are impoverished because they lack basic commonsense constructs such as such as time, space, and physics. Finally, the system does not have to generate explanations *de novo* but rather only needs to chose between already provided answers. Lastly, it does not follow the frame described here of predicting an emotion and providing an explanation for it.

3 Six-Stage Method

The purpose of the method is to acquire knowledge that enables an agent to predict and explain the emotions likely experienced by a given character in a given situation. For example, such knowledge should enable the generation of response to a question such as the following:

> SCENARIO: Tracy wants a banana but gets an apple.
> QUESTION: Explain why she's Happy or Sad?

In this method, an answer to such a question is defined to be an **appraisal**, i.e. an emotion label (such as happy, sad, angry etc) and an explanation justifying the predicted model.

The method is six-phase cyclical process (see Figure 1). It involves participation from laypeople (who function essentially as subject matter experts in the field of commonsense psychology) and modelers or modeling processes that try to maximize fidelity of their models to the input and feedback from the laypeople.

As one iterates through successive cycles, three products are accumulated: (1) appraisals (2) data rating the believability of these appraisals (3) a knowledge base of axioms (or other predictive model that) generate appraisals.

3.1 Phase A: Scenario Generation:

The goal is to generate or obtain scenarios that are expected to cause a character in them to feel an emotion. The modeler needs to decide on a scope appropriate for the current iteration through the cycle - the breadth – the number or range of variation between scenarios – as well as the depth – how complex should the scenarios be. Are the scenarios obtained from some pre-existing source (e.g. a children's story, a novel) or should they be synthetically generated (see Parameterized Questions below).

3.2 Phase B: Humans Generate Appraisals

In order to model human appraisal generation one needs to see actual human generated appraisals. To obtain such data a set of lay people are presented with the scenarios from Phase A and asked to make predictions about what emotion would

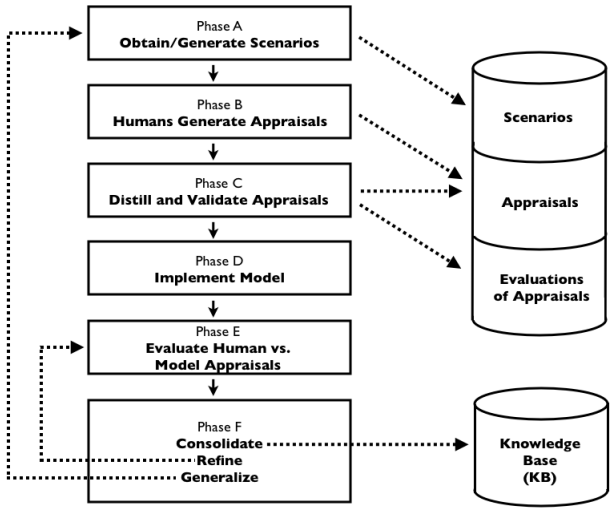

Figure 1: High Level Schematic of the Framework's Six Stages

be felt. Beforehand, the modeler needs to make decisions about scope. How many emotions may they chose from? Does one attempt to control the richness of the explanations via restricting the allowed length or vocabulary?

3.3 Phase C: Distil & Validate Appraisals

The purpose of this stage is to ensure that data from Phase B are simple enough to be modeled yet have good enough fidelity to the original data. If the raw Phase B appraisals need simplification the optional step **C1: Appraisal Distillation** is performed to simplify or narrow the scope of data obtained from Phase B.

In **C2: Validate Appraisals** the purpose is ensure that only high quality items are used as data for modeling in Phase D. A set of human raters independent of the Phase B human appraisers is recruited to evaluate the items on some sort of quality scale. See Figure 5 for an example item.

Ratings are used to remove appraisals from Phase C1 that:

1. are idiosyncratic to a particular appraiser in Phase B
2. have not been corrupted via flawed distillation in C1
3. do not demonstrate falsifiability

Requirements (1) and (2) can be met by removing appraisals that exhibit poor ratings. In severe cases, e.g. flawed distillation, one may need to repeat a prior phase (e.g. C1). Regarding requirement three, recall a key interest is in modeling predictions that are falsifiable. To demonstrate falsifiability, a random fraction of the items seen by a given rater are subject to a manipulation called *reversal* that converts a given item into one that makes the opposite prediction. For example, if the emotion is *Happy* then in the reversed condition it would be Sad. Quality ratings for reversed items should poor and be significantly worse than their unreversed counterparts. If not, they should be filtered out before Phase D.

Due to appraisal diversity as well as the reversal manipulation there multiple items associated with a given scenario. To ensure independent ratings care should be taken that only one item per scenario is assigned to a given rater.

3.4 Phase D: Model Implementation

The modeler (whether an ontologist or machine learning expert) seeks to produce a model that mimics the behavior seen in the appraisals from the results of distillation in Phase C. By model we mean anything with natural language generation capabilities that produces appraisals. The core of the model could be anything from a knowledge base plus inference engine to an algorithm produce via machine learning.

3.5 Phase E: Eval Human vs Model Appraisals

To compare the model to human performance, a new set of humans are presented with a sampling of appraisals – some are human-derived (from Phase C) and others are generated by model (from Phase D). Raters are blind to whether an appraisal was human or model generated. Ratings are used to decide if the model generated appraisals are as believably human as the human generated ones. Reversal is used again to ensure falsifiability of model predictions and guard against chatbot "models".

3.6 Phase F: Consolidate, Refine & Generalize

In light of the Phase E evaluation results, for any given model component, the modeler has the following (non-exclusive) choices:

- **Consolidate**: This is where knowledge acquisition is consummated. Any components that perform sufficiently well in Phase E are added to the knowledge base (KB) of appraisal generator components/axioms that has so far been accumulated. A component is considered satisfactory if it is responsible for appraisals that are: (1) rated in the believable range, (2) are not significantly less believable than their human generated counterparts, (3) are falsifiable (i.e. made significantly unbelievable via the reversal manipulation).
- **Refine**: For any unsatisfactory evaluation results the goal is to understand the deficiencies and refine the model so as to improve them. One then re-runs the model over all the scenarios. For any changes in the generated appraisals, these are fed back to a human evaluators in Phase B.
- **Generalize**: For any components performing satisfactorily, one attempts to apply it to a larger scope of scenarios, cycling back to Phase A. One can find additional sources scenarios, generate them synthetically (see *parameterization* in Phase F). If significant distillation was performed in Phase C, one can loop back to C and relax some of the distillation producing scenarios that are closer to the raw Phase B appraisals.

4 Results of Applying Method

In this section two aims are addressed: (1) testing the hypotheses and assumptions inherent in the method and (2) serving as a "dry run" demonstrating how the method results in knowledge acquisition. In this particular application of the

six staged method, the scope is focused on a particular sub-problem of ATOM termed *goal appraisal* - the appraisal of a scenario with respect to an explicitly given focal goal. Aspects of these results have been described in [Jarrold, 2007] and [Jarrold, 2004] at times in greater depth.

4.1 Apply Phase A: Obtain / Generate Scenarios

As this was the first pass through the method it was important to get a sense as to whether our assumptions about whether a diversity of appraisals would be obtainable at all or whether it might depend on the richness of the given scenario. For this reason two types of scenarios were obtained: simple and complex.

Simple Items were obtained from [Howlin *et al.*, 1999], a clinical workbook of remedial mind-reading exercises for children with autism. It was an especially appropriate source of scenarios because its exercises tap affective mindreading. They should teach and test some of the most important basic or central organizing principles of ATOM because they are aimed at young children. See Figure 2 for an example of a simple item derived from such clinical workbook material.

Figure 2: Example of Simple Item

Scenario Cue :

Goal: Tracy wants a banana.
Outcome: Mommy gives Tracy an apple for lunch.
Question: How will Tracy feel when mummy gives her an apple for lunch - happy, indifferent or sad? Briefly explain why.

Complex Items were based on actual letters to Ann Landers, a popular column one finds in many local newspapers. Typically each letter was about a paragraph in length and written by an adult caught in an emotion-arousing quandary seeking advice from Ann Landers. See Figure 3 for example of an item derived from such a letter.

Figure 3: Example of Complex Item

Scenario Cue :

Goal: The author wants to have a clear conscience about giving an heirloom to the widow of her ex-boyfriend.
Outcome: Many years ago, I dated a man who gave me a family heirloom as a gift. We went our separate ways on very good terms, and he never asked for it back. This gift has always been very special to me because it was made by his grandfather and had been given to his mother. Eventually I married someone else and so did he. I recently learned that this man has died. I decided to give the heirloom to the widow.
Question: How does the author feel when she gives the heirloom to the widow. (pick one of Happy, Indifferent or Sad). Briefly (in 1 to 2 sentences) explain why the person feels that way.

Table 1: Phase B: Tallies of Appraisal Valence

	Simple Items								
	1	2	3	4	5	6	7	8	9
Happy:	9	9	0	1	0	1	9	1	1
Indifferent:	0	0	0	4	8	7	0	3	2
Sad:	0	0	9	4	1	1	0	5	6
	Complex Items								
Happy:	0	2	8	0	0	0	9	0	1
Indifferent:	4	5	0	1	0	2	0	1	0
Sad:	5	2	1	8	9	7	0	8	8

4.2 Apply Phase B: Humans Generate Appraisals

Nine undergraduates were asked to predict and explain whether a character was happy, sad or indifferent to each of 9 simple scenarios and 9 complex ones. Consistent with the generativity hypothesis there was a diversity of emotion labels and explanations for each scenario.

Diversity of Chosen Emotion Labels

One source of appraisal diversity was in seen in the different valences participants chose for each scenario. As can be seen in Table 1, there were many cases in which the same item was appraised in different ways. For example, Simple Item 8 generated a total of 1 *Happy*, 3 *Indifferent*, and 5 *Sad* appraisals.

Diversity of appraisal valence can be quantified via a measure of internal consistency such as Cronbach's Alpha[Kline, 2000]. Typically, educational tests are expected to have high levels of internal consistency therefore high levels of Cronbach's Alpha are desirable for such tests. In our case, diversity of appraisal valence would imply poor consistency (after all there would be no one correct "answer"). Thus low levels of consistency support claims of diversity. The value of this statistic for all of the items was 0.46 which is considered "unacceptable" for an educational test. Such a high level of inconsistency is consistent with our diversity hypothesis. If we restrict analysis to just the simple items, Cronbach's Alpha is 0.56 – in the "poor" range. Restricted to the complex items it is much smaller, -0.69 – in the "unacceptable" range – and thus possibly demonstrating that the complex items exhibit more diversity than the simple items. That said, Cronbach's alpha is typically computed for tests having a large number of items and so these results should be reported with some caution. In view of such a caution, quantitative estimates are consistent with the prediction that there should be many different valences that people attribute to an item. Future work involving more items one should verify if the diversity hypothesis continues to be supported via poor Cronbach's alpha values.

Diversity of Explanations Given

The explanatory diversity is best illustrated in the next section (see 4.3) where one can see that each participants' response was composed of at least one, often more than one "atomic" appraisal. For each scenario, different participants often generated appraisals that contained different atoms or combinations thereof. Due to space considerations quantifying this diversity must be left for future work.

Summing up, in "Apply Phase B" there were two "take aways". First, results were consistent with the diversity hypothesis because for any given scenario there was a variety of predicted emotions and explanations that different participants produced. Second, the actual appraisals generated by the humans in this step comprise a form "acquired knowledge" expressed in natural language that can be passed to the next phase of the method - Distillation and Validation.

4.3 Apply Phase C: Distillation and Validation

C1: Appraisal Distillation

It was judged that the raw appraisals from Phase B were too complex to model directly in a first pass. Thus, distillation/simplification was performed. Specifically, to make the problem of finding and modeling patterns in the appraisals more manageable, the human generated explanations from Phase B were broken into smaller chunks.

As an example, consider the item in Figure 4

Figure 4: Example of Compound Appraisal

Scenario Cue :

Goal: Eric wants to ride in the car.
Outcome: Eric is riding in the train.
Question: How does Eric feel? (Choose from Happy, Sad or Indifferent and explain why he would feel that way?)

Here is how one participant responded to this item.

> *Eric is happy. Why? He maybe* [sic] *excited about the train because he didnt* [sic] *know of his option to ride it and he is still getting transported.*

Via distillation an atomic appraisal of Type II is extracted.

- *Eric is happy. Why? ...he is still getting transported.*

After studying patterns in the atomized data, the following types of appraisal seemed capable of generating a significant proportion of the data:

Three types of appraisal were induced depending on the content of the explanation.

- **Type I:** explicitly indicates focal goal satisfied
- **Type II:** explanation references a substitute goal
- **Type III:** explicitly indicates focal goal not met

Further analysis of Study 1 data suggested that Type I appraisals are almost always assigned the valence of *Happy*, Type II appraisals are often assigned the valence of *Happy* or *Indifferent*, Type III appraisals are often assigned the valence of *Sad* but in rare cases some participants have assigned *Indifferent* as a valence.

The degree to which the assumption that atomic Type I, II and III appraisals can stand alone and be judged of "high quality" by human evaluators is empirically tested in the next validation step in which humans evaluate these distilled appraisals.

Figure 5: Example Item from C2: Humans Rate Appraisals

Example Item A:

Scenario Goal: Toby wants some orange juice.
Scenario Outcome: At bedtime daddy makes Toby some hot chocolate.
Question: How will Toby feel when daddy makes him hot chocolate?
Feeling: Happy
Reason: He got attention from daddy.
Taken together, how believable is the answer given the goal and the reason given?

Highly Unbelievable	Moderately Unbelievable	Neutral	Moderately Believable	Highly Believable
1	2	3	4	5

Table 2: Effect of Reversal Across and Within Conditions

Analysis	Means(sd)		F(df)	p
	Unrev	Rev		
Simple	3.9 (0.5)	2.2 (0.7)	207.85(1,46)	$<.0001$
Complex	3.8 (0.6)	2.3 (0.6)	164.01(1,46)	$<.0001$
Type I	4.7 (0.7)	2.0 (1.2)	42.80(1,36)	$<.0001$
Type II	3.7 (0.8)	2.3 (0.8)	171.6(1,46)	$<.0001$
Type III	3.6 (1.1)	2.18 (1.0)	78.39(1,46)	$<.0001$

C2: Humans Rate Appraisals

The purpose was to validate the atomic appraisal distilled immediately above. First, can the simplified explanations stand on their own as believable appraisals? Falsifiability was tested via the reversal manipulation to verify that only certain valences go with certain appraisal types.

Results (see Table 2) indicate that when one restricts analyses to a single level of Complexity (i.e. Simple or Complex) or of Type (i.e. I, II, or III) one again sees a significant effect for reversal (p ($p < 0.0001$) and that mean believability is greater than 3.0 for unreversed items and less than 3.0 for reversed items. Additionally when one looks within each appraisal type, there is no difference in believability between complex and simple scenarios.

These data support the hypotheses that the distilled atomic Type I, II, and III can stand alone as believable appraisals whether one is dealing with simple or complex scenarios. Additionally, the desired effect of reversal is consistent with the hypothesis that typically the explanations for each appraisal type support falsifiable predictions. Thus, there are reasons to believe that this typology is worth modeling despite being a simplification of the original data. Having been validated the three appraisal types are submitted to Phase D for modeling.

4.4 Apply Phase D: Model Implementation

The objective of this phase is to implement a model based on the appraisal phenomena validated in Phase C. The present implementation occurred via knowledge engineering described in [Jarrold, 2007].

To achieve this the rules validated in Phase C (i.e. those that define/generate Type I, II, and III appraisals) were engineered into a knowledge-based reasoning system known as The Knowledge Machine [Clark and Porter, 1999b] or KM.

In addition, KM's natural language explanation capabilities were leveraged to produce a human readable explanation of each inferred emotion.

Computational Model The model can be separated into three main components: (1) a background knowledge, (2) a scenario representation and (3) an affective mind-reading.

Background Knowledge The component represents general knowledge about everyday objects and events. A simple example involves the assertions that:

```
(superclass Apple Fruit)
(superclass Banana Fruit)
```

Background knowledge like this can be useful in a goal substitution context (e.g., "Tracy wanted an apple but got a banana yet still felt happy because at least she got some fruit."). Ideally, this component should be invariant during successive iterations through this KA loop.

Scenario Representation Component Each scenario cue has the following components: (1) an appraising agent, (2) an overriding goal, (3) an outcome. Each scenario cue is represented as an instance of a special class of data objects in KM that are known as `Situations` - an inference-capable implementation of the situation calculus described in [Clark and Porter, 1999a]. Distinct `Situations` allow one to express mutually contradictory scenario cues.

Affective Mind-reading Component A third knowledge component models affective mind-reading. It contains knowledge about emotions, goals, and appraisal types.

Rules indicate when a Type I, versus a Type II, versus a Type III appraisal applies. For example, one of the Type I model's rule checks to see if the object type desired subsumes the type of the object received. Thereby classifying a scenario such as:

Fred wants an airplane. Fred gets a lear jet.

as a case of Type I because of background knowledge

```
(superclass Learjet Airplane)
```

Type III rules fire when a focal goal is inferably blocked. Type II rules fire when a substitute goal is succesful.

Two types of goal substitution were modeled based on explanation data : *vertical* and *horizontal substitution*. Vertical substitution occurs when generalization of the explicit goal is successful. For example, if Tracy has the goal of *eat apple*, then the goal *eat fruit* can be a vertical substitute. A horizontal substitute goal is any other successful goal that the goal agent can be believably assumed to possess as a background goal. Induced from Phase B data the following background goals were encoded via knowledge engineering in [Jarrold, 2007]:

- *Children want attention from their parents.*
- *Children like to be accompanied by their parents.*
- *People like to eat.*

Generation of Model-Based Appraisals The resulting implementation was used to generate English appraisals (i.e. an inferred emotion and corresponding explanation) for each of the given simple scenarios which in the next phase are validated in comparison with human appraisals from Phase C.

4.5 Apply Phase E: Evaluate Human vs Model Appraisals

Objective

The purpose of this phase is to let a second set of humans blind to whether the appraisal is machine or human generated. As in Phase C2 a portion of the items are subject to valence reversal. This not only verifies the theory-like nature of the model but also identifies empty 'chatbot' like answers which are essentially believable in any reversal context.

Method

To validate the functioning of the model, 130 humans were given items about which to rate believability. Items were varied according to manipulations of two independent variables – *appraisal source* and *reversal*. In the *appraisal-source* manipulation, half of the items presented to the subjects were generated by the computer implementation. The other half were human generated distilled and validated in Phase C.

Note that there was an added twist to the reversal manipulation. Unlike in Phase C2 which had only two levels of reversal unreversed and reversed there was a third more subtle level of reversal - *slightly reversed* in which *Sad -> Indifferent* (rather than *Happy* in the fully reversed condition) and correspondingly, *Happy -> Indifferent* (rather than *Sad*).

Just like in Phase C - Validation, participants rated items believability. One means of achieving the main objective of this phase is to show that the computer model generated believable appraisals that were not significantly less believable than human-generated. Just like in Phase 3b, due to the problematic nature of proving no significant difference, a benchmark significant difference was needed in order to support parity between human- and computer-generated appraisals.

Results and Interpretation

The data indicate that reversal works as predicted. As reported in [Jarrold, 2007] reversal significantly decreased believability whether the appraisal source was the computer model's ($F(2,206) = 85.59$ p ¡ .0001) or the humans' ($F(2,227) = 161.20$ p ¡.0001). Further unreversed items were believable because the mean believabilities from both computer and human appraisers (m(sd) = 3.7(1.2)) versus 4.0 (1.2) respectively) was greater than neutral (i.e., 3.0). Thus, the overall model makes falsifiable predictions. Second, as reported in [Jarrold, 2007] unreversed computer, like human, items of each of the three appraisal types were on average believable (see Table 3). Thus, rating data supports that the model produces reasonably believable appraisals for each type of appraisal.

Although these results were encouraging, unfortunately the Type II model generated appraisals were significantly *less* believable ($p = 0.034$) than their human generated counterparts (see Table 3). This suggests an area where the model needs refinement in Phase F. The results about the strengths and weaknesses of the model are now fed to the final Phase F.

Table 3: Believabilities of Human vs Computer Appraisals

Analysis	Means(sd)		*F(df)*	*p*
Restricted To:	Computer	Human		
Type I	4.8 (0.8)	4.9 (0.2)	1.13(1,64)	0.29
Type II	3.5 (1.2)	3.9 (1.1)	4.74(1,59)	0.034*
Type III	3.5 (1.2)	3.6 (1.3)	0.02(1,39)	0.90

$p < 0.05$.

4.6 Apply Phase F: Consolidate, Refine & Generalize

Results from Phase E help us decide what to do regarding choices between Consolidate, Refine & Generalize for each of the Type I, II and III generators.

Consolidate

Type I, II and III produce appraisals in the believable range and the significant reversal effect demonstrates that each type produces falsifiable predictions. Unreversed Type I and III model-based appraisals were not significantly less believable than their human-based counterparts. On these grounds, the Type I and III could be "consolidated" or considered acquired knowledge. See Refine for what to do about Type II.

Refine

As mentioned above, the model's Type II appraisals were not as believable as their human-generated counterparts. Upon closer inspection one finds that two Type II appraisals had unreversed items in the unbelievable range (e.g. see Figure 4). Interestingly, for both of these items when valence was slightly reversed to *Indifferent*, believability was in the believable range. Furthermore, there were only three cases of vertical substitution. Two of these three had *Indifferent* rated more highly than *Happy*. Therefore, it might be more believable to assume that when vertical goal substitution is involved predict *Indifferent*. Or perhaps substitution of Apple for any Fruit is *Happy* but the other verticals, i.e. any drink for hot chocolate, or any transport for train. These are two proposed model refinement alternatives. They can be independently implemented and the resulting appraisal deltas can be cycled back for evaluation in Phase E and if successful consolidated in Phase F.

Generalize

On these grounds once could consider the existing model as "successful" with respect to the given set of scenarios - it is able to produce quality appraisals and as a whole the model embodies a falsifiable theory. A next step would be to attempt to generalize the model. For example, there is the assumption that vertical goal substitution works. If a child wants some type of fruit will *any* fruit substitute? How high in the generalization hierarchy above fruit will a substitute believably work? To test such hypotheses generation of alternatives via scenario parameterization may be useful. For example an instance like *Tracy wants an banana and gets a banana* might be used as a seed to generate a variety of alternative scenarios via templates such as in Figure 6.

Table 4: Type II Indifferent More Believable than Happy

Goal: Toby wants some orange juice.
Situation: Daddy makes Toby hot chocolate.
Question: Does Toby feel Happy, Sad or Indifferent

Computer Generated Appraisal: *At least he gets to have a drink.*			
Reversal Condition (Emotion)	Unreversed (Happy)	Slightly Reversed (Indifferent)	Strongly Reversed (Sad)
Believability mean(sd)	2.5 (1.3)	3.7 (1.1)	2.8 (1.2)
Human Generated Appraisal: *Because he at least got something.*			
Reversal Condition (Emotion)	Unreversed (Indifferent)	Slightly Reversed (Happy)	Strongly Reversed (Sad)
Believability mean(sd))	3.4 (0.9)	3.3 (1.2)	2.3 (0.6)

Figure 6: Example of Template for Scenario Generation

Scenario Template:

<target-character> wants <object1>. <alt-character> gives <target-character> <object2>

5 Discussion

A six-phase cyclical method has been proposed for acquiring knowledge in the difficult to "pin down" domain of commonsense psychology. A partial piloting of the method on one a single slice of commonsense psychology, i.e. a subproblem of ATOM termed goal appraisal, has tested some of the methods underlying assumptions and resulted in a small amount of knowledge acquisition. In addition something possibly considered a form of implicit knowledge acquisition has occurred: the creation of a modest corpus of scenarios, appraisals and appraisal ratings.

Results are consistent with the generativity hypothesis. For both simple and complex scenarios, people naturally generate and endorse a variety of mutually inconsistent appraisals and that these constitute falsifiable predictions. Focusing on just the simple scenarios a rule-based model of appraisal phenomena was developed and validated via human raters. Rating data provided feedback on where the model needs the most refinement (i.e. Type II).

One challenge for really scaling this approach up is the lack of a general (i.e. non-affective) common sense knowlege. A partial workaround is to narrow the scope of scenarios and appraisals according to whatever a given common sense knowledge system has.

There are some concrete next steps for strengthening the case for this method. Foremost among these is the question of reducing the level of human effort required for certain phases. One focused area to look at is the distillation step of Phase C.

Perhaps the effort required can be reduced simply by restricting the allowable length and vocabulary of explanations in Phase B? Perhaps NLP techniques can be used to automatically extract atomic appraisals.

Another area for potential (semi) automation is Model implementation (Phase D) and Refinement (in Phase F). The slow rate of knowledge engineering is well known. A variety of (semi-)automated approaches are relevant background. KNOWBOT [Hixon *et al.*, 2015] is a system that learns to refine a knowledge graph via natural language dialog with the user. Additional approaches to remedying the challenges of model implementation include rule induction, statistical NLP or other machine learning approaches. See [Bader *et al.*, 2008] for an example of a RNN that learns to reason with first order logic. See [Weston *et al.*, 2015] for a system that learns to answer questions about simple scenarios and [Bowman *et al.*, 2016] for learning textual entailment. Such methods may entirely replace a traditional rule-based approach or be deployable in a hybrid fashion. For a hybrid example, word embeddings may replace a subsumption hierarchy in recognize cases of goal substitution providing evidence that, say, a generic fruit is a good enough substitute for an apple based on semantic distance.

Virtually all of these automated methods require vast amounts of training data which points to an added benefit of this process beyond acquisition of formal knowledge. The scenarios, the appraisal data and appraisal evaluation data are also useful for training and testing of such systems.

Scaling this method will require a large multi-team effort. It is for this reason we seek feedback and collaborators.

References

[Adams *et al.*, 2016] Sam S. Adams, Guruduth Banavar, and Murray Campbell. I-athlon: Towards a multidimensional turing test. *AI Magazine*, 37(1):78–84, 2016.

[Bader *et al.*, 2008] Sebastian Bader, Pascal Hitzler, and Steffen Hölldobler. Connectionist model generation: A first-order approach. *Neurocomputing*, 71(13–15):2420–2432, 2008.

[Bowman *et al.*, 2016] Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. A fast unified model for parsing and sentence understanding. *CoRR*, abs/1603.06021, 2016.

[Clark and Porter, 1999a] P. Clark and B. Porter. Km - situations, simulations and possible worlds. Retrieved April 4, 2001 from `http://www.cs.utexas.edu/users/mfkb/RKF/situations.html`, 1999.

[Clark and Porter, 1999b] P. Clark and B. Porter. The knowledge machine 1.4: Users manual. Retrieved April 4, 2001 from `http://www.cs.utexas.edu/users/mfkb/RKF/km.html`, 1999.

[Gordon, 2016] Andrew S. Gordon. Commonsense interpretation of triangle behavior. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, Phoenix, Arizona, 2016.

[Hayes, 1985] P. J. Hayes. The second naive physics manifesto. In J. R. Hobbs and R. C. Moore, editors, *Formal Theories of the Commonsense World*, pages 1–36. Ablex, Norwood, NJ, 1985.

[Heider and Simmel, 1944] F. Heider and M Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259, 1944.

[Hixon *et al.*, 2015] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *Proceedings of the the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA*, 2015.

[Howlin *et al.*, 1999] P. Howlin, S. Baron-Cohen, and J. Hadwin. *Teaching children with autism to mind-read: a practical guide for teachers and parents*. J. Wiley & Sons, Chichester, NY, 1999.

[Jarrold and Yeh, 2016] W. Jarrold and P.Z. Yeh. The Social-Emotional Turing Challenge. *AI Magazine*, 37(1):31–38, 2016.

[Jarrold, 2004] William Lawrence Jarrold. *Towards a theory of affective mind: computationally modeling the generativity of goal appraisal*. The University of Texas at Austin, 2004.

[Jarrold, 2007] William Jarrold. Treating autism with the help of artificial intelligence: A value proposition http://www.ai.sri.com/pubs/files/1518.pdf. In *Workshop on Agent-Based Systems for Human Learning and Entertainment (ABSHLE)AAMAS ACM 2007*, 2007.

[Kline, 2000] P. Kline. *The handbook of psychological testing (2nd ed.)*. Routledge, London, 2000.

[Levesque, 2011] H. J. Levesque. The Winograd Schema Challenge. In *Proc. of the CommonSense-11 Symposium, March 2011*, 2011.

[Maslan *et al.*, 2015] Nicole Maslan, Melissa Roemmele, and Andrew S. Gordon. One Hundred Challenge Problems for Logical Formalizations of Commonsense Psychology. In *Proceedings of the Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense-2015)*, Stanford, CA, March 2015.

[Ortony *et al.*, 1988] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotion*. Cambridge University Press, Cambridge, UK, 1988.

[Ortony, 2001] A. Ortony. On making believable emotional agents believable. In R. Trappl, P. Petta, and S. Payr, editors, *Emotions in humans and artifacts*, pages 189–213. MIT Press, Cambridge, MA, 2001.

[Ravenscroft, 2010] Ian Ravenscroft. Folk psychology as a theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab: Center for the Study of Language and Information, Stanford University, fall 2010 edition, 2010.

[Weston *et al.*, 2015] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698, 2015.