Advances in Cognitive Systems 7 (2018) 23–38 Submitted 6/2018; published 12/2018

Interpretation of the Heider-Simmel Film Using Incremental Etcetera Abduction

Andrew S. Gordon GORDON@ICT.USC.EDU
Institute for Creative Technologies, University of Southern California, Los Angeles, CA 90094 USA

Abstract

In 1944, psychologists Fritz Heider and Marianne Simmel created a short, 90-second animated film depicting two triangles and a circle moving around a box with a hinged opening, and reported how subjects viewing the film anthropomorphized the three shapes as characters with humanlike goals, emotions, and social relationships. In this paper we model this type of high-level reasoning as a process of probability-ordered logical abduction (Etcetera Abduction), where the interpretation of the film is incrementally constructed by disambiguating observed movements in the contexts of multiple running hypotheses. We describe a target interpretation and knowledge base that we used in a series of experiments to investigate the effects of varying the window size and number of running hypotheses maintained during the interpretation.

1. Introduction

In their early investigations of the psychology of perception, Fritz Heider and Marianne Simmel created a short, 90-second animated film[1] depicting two triangles and a circle moving around a box with a hinged opening (Heider & Simmel, 1944). After viewing the film, experimental subjects were asked to report what they had seen in the movie, to which they responded with narratives of the three shapes as characters with humanlike goals, emotions, and social relationships. As Heider (1958) would later propose in his influential book, *The Psychology of Interpersonal Relations*, viewers turned to anthropomorphic interpretations of the shapes' movements, applying commonsense theories of human psychology to explain behavior when physics-based explanations fail. This human tendency to adapt an *intentional stance* (Dennett, 1989) is seen as a foundation for human social cognition, and likewise has been a fundamental concern in successful human-computer interaction.

Modeling this human cognitive ability in software systems poses a number of difficult challenges, owing to the richness of human commonsense psychological theories and the subjective nature of social interpretation problems. In previous work (Gordon, 2016) we made progress by using a 100-question evaluation set called TriangleCOPA (Maslan et al., 2015), where each question presents the system with a (formal) description of a short interaction between characters in the original Heider-Simmel setting, and asks which of two interpretations would be preferred as more plausible by human raters. We modeled the interpretation task as a problem of logical abduction, analogous to Hobbs et al.'s (1993) formulation of language interpretation as abduction, where the

1. The film is viewable online at various websites, including `https://www.youtube.com/watch?v=n9TWwG4SFWQ`

© 2018 Cognitive Systems Foundation. All rights reserved.

system chooses the interpretation that can be logically entailed by assumptions with the highest joint probability, given a knowledge base of probabilistic axioms.

Although promising, our previous solution also underscored the problem of scalability. Given a reasonably large knowledge base, a problem with nine character actions proved to be intractable, owing to the strategy of combinatorial search used in logical abduction. Although subsequent work has shown that advanced optimization techniques can be applied to this approach (Inoue & Gordon, 2017), narratives as large as the original Heider-Simmel film – depicting 75 or more character actions – would still be intractable to interpret in this manner. Considering how people watch a movie like the 90-second Heider-Simmel film or a 120-minute Hollywood feature, we see instead a robust capacity for real-time, incremental interpretation. Instead of waiting until the end of the movie to begin processing what they saw, people's running interpretations of the events, as they unfold, are evidenced by their laughs, groans, and other reactions as audience members.

In this paper, we explore an approach to incremental interpretation of ordered events, specifically a formalization of the character actions evident in the original Heider-Simmel film. To support the development and evaluation of our model of incremental interpretation, we authored a single target interpretation of this film along with the specific set of axioms that would produce this interpretation using our original proposal for Etcetera Abduction, if computationally tractable. We describe a new algorithm, Incremental Etcetera Abduction, that segments ordered observations into small sets that are incrementally interpreted in the context of multiple running interpretations, producing as output a final set assumptions that logically entails the entire larger set. As our approach is not guaranteed to find the globally optimal (most probable) solution, we investigate how segment size and interpretation count effect optimality.

2. Related Work

People's interpretations of the Heider-Simmel film has been extensively researched in the area of perceptual psychology over many decades (Shor, 1957; Greenberg & Strickland, 1973; Massad et al., 1979). In the computational modeling of this interpretation processes, the most comprehensive work is that of Thibadeau (1986). Here the position of shapes in each of 1,690 frames in the original Heider-Simmel film are annotated by hand using formal descriptions, and differences in descriptions across frames are used to derive action descriptions using formal, hierarchically-organized action schemas. Thibadeau then compares the times of derived action descriptions with empirical data of perceived actions, collected in Massad et al.'s (1979) early work where 55 students clicked a button whenever they perceived an action occurred. Finding high correlation across action categories, Thibadeau argues that clarity of intention is a better predictor of action perception than their participation in high-level plan structures attributed to characters.

In our present research, we encoded the original Heider-Simmel film manually in terms of event descriptions that are roughly equivalent to Thibadeau's collection of action schemas. We treat these descriptions as the given observations in the interpretation task, where the assumed plans, goals, emotions, and social relationships among characters explain these observations as implication-rich intentional actions. Although Thibadeau's formal approach to low-level action perception and segmentation was well-suited to this particular film, we believe that contemporary machine-learning

approaches to action perception (e.g., Roemmele et al., 2016) will ultimately prove more robust in future perception-interpretation pipelines. In any case, it is the output of the action perception process investigated by Thibadeau that is the input to the higher-level interpretation process investigated in our current research.

Incremental interpretation has been previously explored in systems that incrementally maintain explanations of actions and events over time. The DiscoverHistory system (Molineaux & Aha, 2015) takes an agent-based approach to infer the sequences of events and assumptions that explain a series of agent observations, incrementally improving inconsistent explanations through refinement. The UMBRA system (Meadows et al., 2013) performs plan understanding by incrementally constructing explanations using a knowledge base of hierarchical task networks and domain knowledge, incrementally maintaining the smallest set of assumptions that explain the observed actions of an agent. Our present work differs most from DiscoverHistory and UMBRA in that multiple running explanations are maintained during incremental interpretation, all of which logically entail the observations without inconsistencies, and which are ordered by the joint probability of assumptions rather than a heuristic cost function or bias toward smaller sets of assumptions.

Narrative interpretation as pursued in our present work is closely related to research on plan recognition, which also aims to ascribe intentions to agents that explain their observable behavior (Kautz & Allen, 1986). Whereas we pursue “bottom-up" algorithms that construct explanations of observations by backward-chaining through a knowledge base, the more typical strategy in plan recognition takes a “top-down" approach, constructing explanations from a library of plans (hierarchical task networks) and maintaining a probability distribution over the space of candidates. Our current work on incremental interpretation is most similar to online plan recognition, where algorithms incrementally maintain hypotheses about an agent’s plans given an ordered set of observations. Geib and Goldman (2009) present the PHATT algorithm for online plan recognition, a “top-down" probabilistic algorithm that considers all possible plans given the observations, maintaining all possible explanations for future unseen agent actions.

As with other abductive reasoning problems, the elaboration of all possible explanations is intractable beyond small plan libraries and observation sequences, leading other researchers to develop techniques for ordering and pruning the space of possible plan explanations. The DOPLAR algorithm (Kabanza et al., 2013) expands hypotheses only for those plans that are most probable, capitalizing on fast computation of upper and lower bounds on goal probabilities to avoid the costly computation of exact probabilities for candidate hypotheses. The CRADLE algorithm (Mirsky et al., 2017) incrementally prunes the space of candidate hypotheses through the application of several domain-independent filters, retaining plans that have recently been extended, that make fewer commitments about future observations, that are more compact than other explanations, and that have a higher probability of generating the observation sequence. Our present work is similar to these approaches in that the probabilities of hypotheses are used to select those that are maintained as running interpretations, thereby avoiding the exhaustive exploration of an intractably large space of candidate explanations.

Although some parallels can be drawn between the knowledge bases used in our present work and the plan libraries used in plan recognition research, our work differs most in its inclusion of world knowledge beyond plans in the explanations of observations, and in its use of first-order logic

to represent this knowledge. Our approach is strictly “bottom-up" in the generation of explanations from observations, and makes no special consideration for plans as the best (or only) explanations of observed behavior. We believe our present work may be applicable to future investigations of plan recognition research that are more broadly scoped to consider multi-agent behaviors, context, exogenous events, agent traits, and non-causal factors in behavior explanations.

3. Target Interpretation

To support the development and evaluation of our model of incremental interpretation, we authored a formal representation of the sequence of actions that are observable in the original Heider-Simmel film, along with single interpretation to serve as a target for the reasoning process. Although previous studies with the Heider-Simmel film have demonstrated remarkable agreement about how the film is interpreted by experimental subjects (Shor, 1957), our aim was not to define the definitive interpretation, but rather to restrict the overall problem from one of open-ended interpretation to one that was primarily concerned with scale.

Using the video annotation software ELAN (Wittenburg et al., 2006), we began by identifying spans within the original Heider-Simmel film where any of the three characters were moving, and devising a small vocabulary of action descriptions that best characterized the objective (without interpretation) observable action. In actuality, some degree of subjective interpretation is needed to segment and categorize the actions in this film, but our aim was to devise a small vocabulary of actions that could completely describe the film in qualitative terms. This effort produced 22 labels for a total of 76 actions observed in the film.

We then divided the 90-second film into 11 segments, each comprised of a small set of highly-related actions, according to our own interpretation of this film. For each segment, we introspectively considered what assumptions about the characters and their relationships best explained the observations in each segment, e.g., the nodding of the big triangle toward the little triangle was best explained by its disapproval for attempting to defend against the big triangle’s attacks, and the encircling of the the little circle by the little triangle is best explained by the shared joy they felt in their escape from the big triangle. Although our interpretation was subjective, it closely mirrored that which was reported by Heider and Simmel (1944) as representative of the interpretation commonly made among their experimental subjects.

Finally, we encoded both the observed actions and our assumptions for each segment as sets of literals in first-order logic using the Common Logic Interchange Format (International Organization for Standardization, 2007), shown in Table 1. Here we employed the same event notation used in previous work to solve the TriangleCOPA problem set (Gordon, 2016). This notation affords the easy expression of second-order relations by reifying predications as their own first arguments, e.g., the event of shielding oneself from an attack can use the event of the attack as an argument of its own reification (segment 3). In these representations, the big triangle, little triangle, and circle are represented using the constants `BT`, `LT`, and `C`, respectively, with additional constants for the door (`D`) and the walls of the box (`W1-W3`). Reified events for observations are represented as numbered constants (`E2-E77`), while events for assumed literals in this target interpretation are represented as variables preceded by a question mark, e.g. `?a3` in segment 3. Repeated actions are listed in Table 1 as single literals, using shorthand for the event constants, e.g., `E11-E13` in segment 3.

Table 1. Observations (o) and target interpretation (t) of the eleven segments in the Heider-Simmel film.

1. *Arrival:* LT and BT arrive and BT exits the box, because it is investigating a sound.
   o: `(arrive' E2 LT)(arrive' E3 C)(open' E4 BT)(exit' E5 BT)`
   t: `(hear' ?a1 BT ?a2)(investigate' ?b1 BT)`
2. *Scuffle:* BT pushes LT to attack. LT hits BT to defend. BT nods in disapproval.
   o: `(push' E6 BT LT)(hit' E7-9 LT BT)(nod' E10 BT LT)`
   t: `(attack' ?a2 BT LT)(defend' ?b2 LT BT)(disapprove' ?c2 BT LT)`
3. *Beating:* BT pushes LT to attack. LT hits BT to defend. C half-closes the door as shelter.
   o: `(push' E11-E13 BT LT)(hit' E14 LT BT)(push' E15-17 BT LT)(halfclose' E18 C)`
   t: `(attack' ?a3 BT LT)(defend' ?b3 LT BT)(shield' ?c3 C ?a3)`
4. *Scolding:* BT nods at LT in disapproval of its defense. BT hits and misses LT to attack.
   o: `(nod' E19 BT LT)(hit' E20-E21 BT LT)(miss' E22-E25 BT LT)(nod' E26 BT LT)`
   t: `(disapprove' ?a4 BT LT)(defend' ?b4 LT BT)(attack' ?c4 BT LT)`
5. *Flinch:* BT approaches C, who flinches and enters the box, and closes the door for shelter.
   o: `(approach' E27 BT C)(flinch' E28 C)(enter' E29 C)(close' E30 C)`
   t: `(attack' ?a5 BT C)(shield' ?b5 C ?a5)`
6. *Entrapment:* BT opens the door, enters, and closes door to attack C, while LT scales the wall.
   o: `(open' E31 BT)(enter' E32 BT)(close' E33 BT)(scale' E34 LT)`
   t: `(attack' ?a6 BT C)`
7. *Missing:* C meanders and shakes in fear. BT shuffles and misses C. Concerned, LT opens door.
   o: `(meander' E35 C)(miss' E36 BT C)(shuffle' E37 BT)(shake' E38 C)`
      `(miss' E39-E43 BT C)(halfopen' E44 LT)`
   t: `(attack' ?a7 BT C)(fear' ?b7 C)(concern' ?c7 LT)`
8. *Escape:* C exits as LT closes door. BT pushes the jammed door. LT and C escape in joy.
   o: `(exit' E45 C)(close' E46 LT)(hit' E47-E49 BT D)(push' E50 BT D)`
      `(touch' E51-E54 LT C)(encircle' E55 LT C)(touch' E56 LT C)`
   t: `(jammed ?a8 D)(sharedjoy' ?b8 LT C)(escape' ?c8 C)`
9. *Chase:* BT opens the door and circles around to attack LT, who departs to escape.
   o: `(open' E57 BT)(exit' E58 BT)(circlearound' E59,E65 LT)`
      `(circlearound' E60,E66 C)(circlearound' E61,E64 BT)(enter' E62 BT)`
      `(exit' E63 BT)(depart' E67 LT)(depart' E68 C)`
   t: `(attack' ?a9 BT LT)(escape' ?b9 LT)`
10. *Rage:* BT spins around, then opens and closes the door in rage.
    o: `(spin' E69 BT)(close' E70 BT)(open' E71 BT)(close' E72 BT)`
    t: `(rage' ?a10 BT)`
11. *Destruction:* BT busts the box, and pushes its walls, in rage.
    o: `(bust' E73 BT)(bust' E74 BT)(push' E75 BT W1)(push' E76 BT W2)`
       `(push' E77 BT W3)`
    t: `(rage' ?a11 BT)`

4. Knowledge Base for the Target Interpretation

Our approach was to first author a single knowledge base of axioms sufficient to find the target interpretation for each segment individually, using the original algorithm for Etcetera Abduction (Gordon, 2016), then devise a new incremental interpretation algorithm capable of replicating this interpretation given all observations at once.

Etcetera Abduction is a logic-based reasoning method that searches for sets of literals that, if true, would logically entail a set of input literals given a knowledge base of definite clauses in first-order logic. The reference implementation executes this search by first identifying all possible sets of entailing assumptions for each input literal, independently, and then composing solutions for all input literals by taking the Cartesian product of these sets. Where literals in these composite solutions can be logically unified, additional solutions are created with the appropriate variable substitutions when necessary.

Etcetera Abduction differs from other abductive reasoning methods in that all axioms in the knowledge base (definite clauses) include a literal in the antecedent known as the “etcetera literal," unique to the individual axiom, that reifies the additional, unspecified facts that must also be true to guarantee that the first-order axiom always holds. Originally proposed by Hobbs et al (1993), building on McCarthy's use of abnormality literals as a means of handling defeasibility in first-order logic (McCarthy, 1986), etcetera literals are used in Etcetera Abduction to encode the conditional probability of the consequent in a definite clause given the remaining literals in the antecedent, or the consequent's prior probability in the case that the etcetera literal is the only literal in the antecedent. Here, we follow the convention of encoding these probabilities as real number constants in the first argument position of each etcetera literal.

Etcetera Abduction requires that each axiom contains a unique etcetera literal, and that backchaining on input literals always terminates at etcetera literals representing prior probabilities. When identifying all possible sets of entailing assumptions for a given input literal, Etcetera Abduction only considers assumption sets comprised entirely of etcetera literals, limited by a parameter that limits the depth of backchaining. The joint probability of solutions can then be computed as the product of each etcetera literal's probability. In this way, each unification that is made across assumptions for different input literals identifies a common factor, both figuratively and literally, increasing the probability of the solution. Exhaustively searching all possible combinations and unifications, Etcetera Abduction identifies the most probable set of assumptions (etcetera literals) that logically entail the given input literals.

Working on each segment independently, we created only the axioms necessary so that the most probable set of assumptions entailed the target interpretation. For example, segment 2 in Table 1 depicts five character actions (a nod, three hits, and a push) that are explained by the disapproval of the big triangle over the small triangle's defense against its attack. For this segment, we created the six axioms in Table 2, in Common Logic Interchange Format, so that the three target literals are also entailed by the most probable set of etcetera literals that also entail the five observations.

Figure 1 depicts the entailing relationships between the assumptions (ovals), the target literals (single-line rectangles), and the given observations (double-line rectangles) for segment 2, as identified in the most probable solution. In all, 81 definite clauses were authored to produce the target

Table 2. Six axioms authored to produce the target interpretation of segment 2.

1. Why nod? Maybe disapproval
```
(if (and (disapprove' ?e1 ?x ?y)(etc1_nod 0.1 ?e ?e1 ?x ?y))
  (nod' ?e ?x ?y))
```
2. Why disapprove? Maybe against defend
```
(if (and (defend' ?e1 ?y ?x)(etc1_disapprove 0.5 ?e ?x ?y ?e1))
  (disapprove' ?e ?x ?y))
```
3. Why hit? Maybe to defend
```
(if (and (defend' ?e1 ?x ?y)(etc1_hit 0.1 ?e ?e1 ?x ?y))
  (hit' ?e ?x ?y))
```
4. Why defend? Maybe someone is attacking
```
(if (and (attack' ?e1 ?y ?x)(etc1_defend 0.5 ?e ?e1 ?x ?y))
  (defend' ?e ?x ?y))
```
5. Why push? Maybe to attack
```
(if (and (attack' ?e1 ?x ?y)(etc1_push 0.1 ?e ?e1 ?x ?y))
  (push' ?e ?x ?y))
```
6. Why attack? Attacks sometimes happen (prior probability)
```
(if (etc0_attack 0.1 ?e ?x ?y)
  (attack' ?e ?x ?y))
```

interpretations for each of the 11 segments, of which 37 encoded prior probabilities. In all axioms, the numerical probabilities were selected without regard to empirical data.

5. Incremental Etcetera Abduction

At a high-level, our approach to incremental abduction was to break up large sequences of observations into smaller segments of size $window$, and sequentially interpret each segment in the context of solutions found for all previous segments. From this perspective, the important considerations are how previous solutions are represented, and how they influence the interpretation of the current segment. A guiding principle in our approach was to preserve the central tenet of logical abduction that the resulting solutions logically entail the entire sequence observations. Accordingly, we represent a $context$ as a conjunction of etcetera literals that logically entail all literals observed in previous segments, the most probable of which are maintained as a set of running interpretations of size $beam$. After the last segment in a large problem is processed, this set represents the most probable set of assumptions that logically entail all observations found for given values of $window$ and $beam$.

We explored several mechanisms by which the beam of contexts might influence the interpretation of the current segment. To find globally optimal solutions, it is necessary to unify the entailing assumptions of the current segment with those represented in the contexts. However, our early at-

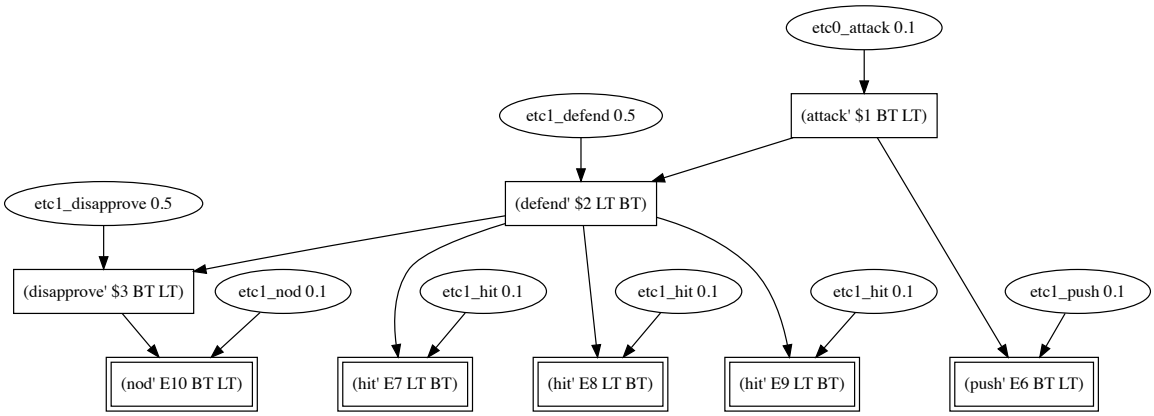

Figure 1. The target interpretation of segment 2, depicting the entailment relationships between the assumptions (ovals), the target interpretation (single-line rectangles), and observables (double-line rectangles).

tempts to devise an algorithm suffered from combinatorial explosions in the number of unifications to be considered, which worsened as the contexts grew in size with each subsequent segment.

Our solution to this problem was to check for possible context unifications as early as possible, *before* full solutions for the current segment are composed. Our approach was to modify the backward chaining algorithm used in our original implementation of Etcetera Abduction. In its original form, this algorithm identifies all conjunctions of etcetera literals that logically entail a single observation by backchaining on knowledge base axioms (definite clauses) to a given depth. Etcetera Abduction subsequently composes full solutions from the Cartesian product of the sets for each input observation. We modified this backchaining algorithm to accept a $context$ as an additional parameter, and allowed the algorithm to drop assumptions from output conjunctions when they could be unified with literals in the context. The resulting algorithm identifies all conjunctions of etcetera literals etc where $etc \wedge context$ logically entail the input observation.

Table 3 lists pseudocode for our approach to Incremental Etcetera Abduction, showing the two primary functions INCREMENTAL and ETCABDUCTION*. The main function, INCREMENTAL, maintains a $beam$ of running $contexts$ that logically entail all previous segments. The function proceeds by processing input literals obs in segments of size $window$ until none remain, then returning these contexts as solutions. Entailing assumptions for each segment are found independently for each context, via the function ETCABDUCTION*. Variables in these solutions are replaced with Skolem constants (via the function SKOLMEIZE), then combined with its context to form a solution for all segments processed thus far. These combinations are sorted by their joint probability, computed as the product of probabilities encoded in the etcetera literals, and the most probable subset becomes the new $beam$ of $contexts$.

The supporting function, ETCETERAABDUCTION*, is identical to the original (non-incremental) formulation of Etcetera Abduction except for the addition of a $context$ parameter, representing assumptions (a conjunction of etcetera literals) that logically entail previously-processed segments. This parameter is passed to the supporting function BACKWARDCHAIN*, described above, which modifies the original version to return all conjunctions of etcetera literals etc found by backchaining

Table 3. Incremental Etcetera Abduction.

```
1: function INCREMENTAL(obs, kb, depth, window, beam)
2:     contexts ← {∅}
3:     while obs ≠ ∅ do
4:         current ← POP_n(obs, window)
5:         combinations ← ∅
6:         for each context in contexts do
7:             solutions ← ETCABDUCTION*(current, kb, depth, context)
8:             for each solution in solutions do
9:                 solution ← SKOLEMIZE(solution)
10:                combination ← APPEND(solution, context)
11:                PUSH(combination, combinations)
12:        ordered ← SORTBY(combinations, JOINTPROBABILITY)
13:        contexts ← POP_n(ordered, beam)
14:    return contexts
15: function ETCABDUCTION*(obs, kb, depth, context)
16:    setOfSets ← ∅
17:    solutions ← ∅
18:    for each observation in obs do
19:        EtcAntecedents ← BACKWARDCHAIN*(observation, kb, depth, context)
20:        PUSH(setOfSets, etcAntecedents)
21:    for each conjunction in CARTESIANPRODUCT(setOfSets) do
22:        combinations ← CRUNCH(conjunction)
23:        PUSH_n(solutions, combinations)
24:    return SORTBY(solutions, JOINTPROBABILITY)
```

in kb from $observation$ to $depth$ such that $etcs \wedge context$ entails $observation$. Unmodified is the supporting function CRUNCH, which returns the set of all conjunctions resulting from the powerset of substitutions for unifiable literals in an input conjunction.

By adjusting the parameters of $beam$, $window$, and $depth$, our approach to Incremental Etcetera Abduction affords some ability to tackle very large interpretation problems with constrained computational resources, albeit without the guarantee of finding the globally optimal solution. In this approach to incremental abduction, solutions with higher probability can be missed when their evidence is distributed across multiple segments such that the common factors are initially seen as improbable, and dropped from an insufficiently large beam before they can be used in the proof of latter observations. Accordingly, interpretation problems where such evidence is relatively close should benefit from larger window sizes, so as to encourage their interpretation within the same segment. Likewise, interpretation problems where such evidence is very distant in the input sequence may benefit from a larger beam, so that initially improbable solutions might remain on the beam until latter evidence is processed.

From the pseudocode, it can be seen that setting the $window$ parameter to the length of input observations obs or greater makes INCREMENTAL functionally equivalent to the original formu-

lation of Etcetera Abduction. In such case, the entire input would be processed as $current$ by ETCABDUCTION* with $context$ parameter set to the null context $\emptyset$, with the resulting solutions sorted (twice!) by their joint probability and returned as $contexts$.

In contrast, an infinite $beam$ setting can still fail to find solutions identified by the original formulation of Etcetera Abduction. In many interpretation problems there will be common factors (unifications) across segments, and our approach can find them in all cases (given sufficient $beam$), *except* in one special case. An assumption (etcetera literal) for an earlier segment containing a universally quantified variable will fail to unify with an assumption for a latter segment with a constant in the same argument position. The reason is that variables in solutions are replaced with Skolem constants before added to contexts (line 9 in Table 3), and literals with different constants in the same position fail to unify following the rules of logical unification. The introduction of Skolem constants in our approach greatly simplifies the representation of contexts, allowing them to remain fixed throughout the analysis of a single segment, but introduces this shortcoming that remains to be addressed in future work.

We implemented our approach to Incremental Etcetera Abduction in Python 3 by modifying the original reference implementation of Etcetera Abduction, and added this new functionality to its open-source repository.[2]

6. Interpretation of the Heider-Simmel Film

Using our Python implementation of Incremental Etcetera Abduction, we generated an interpretation of the Heider-Simmel film using our knowledge base of axioms and formalization of the event sequence as input. For this one interpretation, we set $depth$ to 5, as used in developing the knowledge base for the target interpretation, and set the $window$ to 5 and the $beam$ to 10.

Figure 2 depicts the graphical proof structure of the most probable solution. While the full graph in Figure 2 is admittedly too small to inspect in detail, the overall narrative structure of the Heider-Simmel film is evident, with the first two-thirds of the story interpreted as attacks by the big triangle on the little triangle and circle, leading to a dramatic escape that prompts the big triangle to destroy the box in a fit of rage.

The four magnified insets in the figure illustrate some of the key facets of the incremental interpretation. Inset (a) shows that our incremental approach is able to replicate the target interpretation depicted in Figure 1, where the nods of the big triangle toward the little triangle are seen as expressing disapproval for the little triangle's defense against the big triangle's attacks. Inset (b) shows that the big triangle's aim of attacking the little triangle is a common factor in the interpretation of the entire first half of the film. Here the circle entering and half-closing the door of the box is seen as an act of shielding it from this attack, with a Skolem constant reifying the attack appearing as an argument of the shielding event. Inset (c) shows the interpretation of a moment late in the film when the little triangle touches and moves around the circle, explained by a feeling of shared joy with the circle owing to its escape from the big triangle's attacks.

Inset (d) shows the interpretation of the film's ending where the big triangle's rage is seen as the explanation for closing doors, busting walls, and pushing walls around. The leftmost literal in

2. The software is available at `https://github.com/asgordon/EtcAbductionPy`.

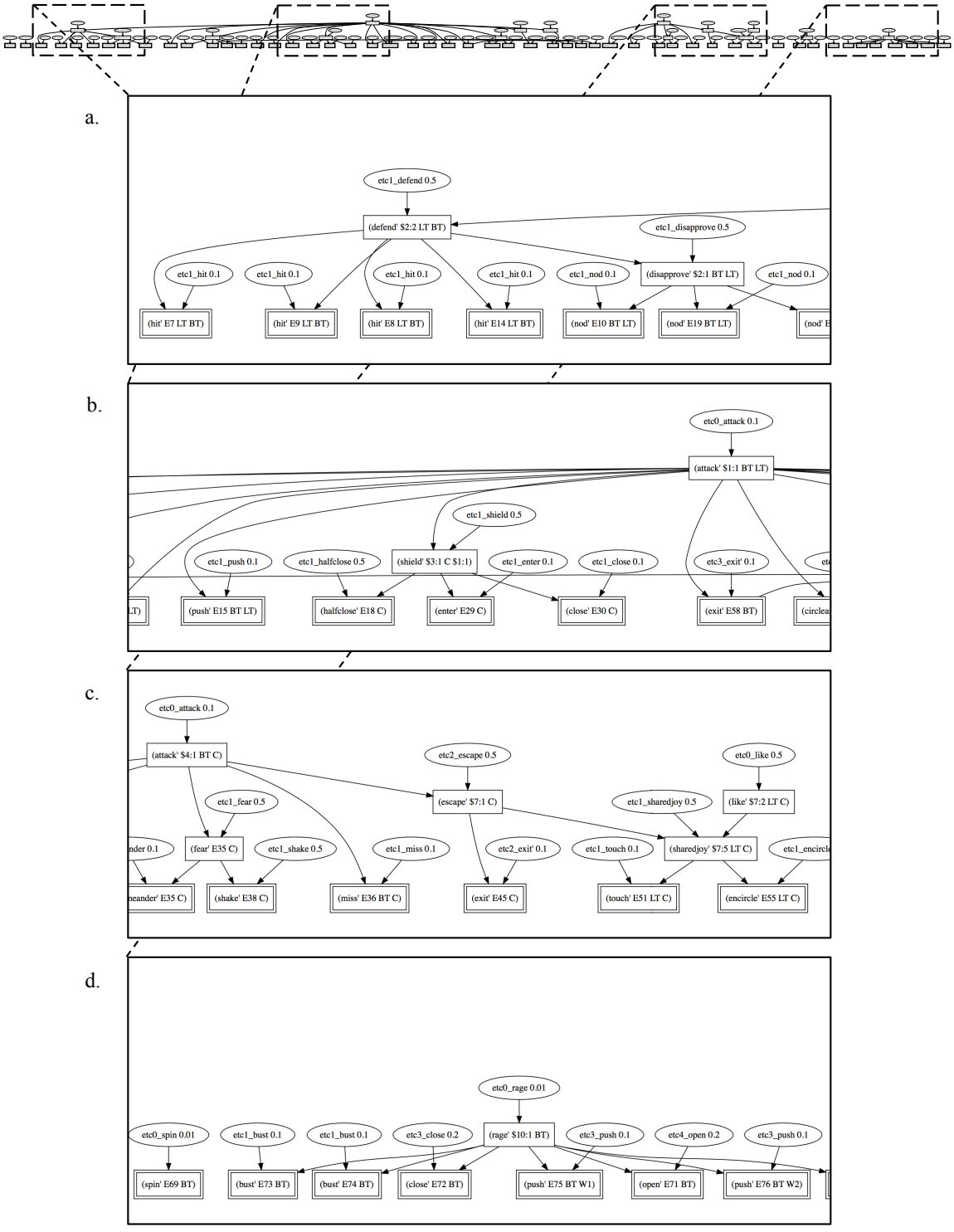

Figure 2. A generated interpretation of the Heider-Simmel film (window=5, beam=10).

Table 4. Precision and recall at different beam and window settings (*T.O.* = timed out at 600 seconds).

	Beam							
	1		10		100		1000	
Window	*prec.*	*rec.*	*prec.*	*rec.*	*prec.*	*rec.*	*prec.*	*rec.*
1	.55	.60	.55	.60	.52	.58	*T.O.*	
2	.61	.65	.67	.70	.67	.70	.67	.70
3	.70	.74	.71	.75	.71	.75	.73	.74
4	.75	**.79**	.75	**.79**	*T.O.*		*T.O.*	
5	**.79**	**.79**	**.79**	**.79**	*T.O.*		*T.O.*	
6	*T.O.*		*T.O.*		*T.O.*		*T.O.*	

this inset, the big triangle spinning around, is unconnected from the larger interpretation, explained only by its prior probability. In our target interpretation for this event, however, the big triangle is spinning due to its rage – the same rage that explains subsequent events. Our approach fails to make this connection due to an unfortunate segmentation of the event sequence that separates the spinning from latter evidence of rage. Seeing the spin in insolation, the knowledge base favors the prior probability over an interpretation involving the big triangle's rage, and this latter interpretation is dropped off of the beam before it can be promoted by further evidence.

7. Tuning the Window and Beam Parameters

We evaluated whether our approach could generate interpretations closer to the target interpretation by varying the parameters of $beam$ and $window$. Our hypothesis was that using a very large beam would yield performance as good as when using a large window, given the same constraints on computational resources. We reasoned that high values in either parameter allow the algorithm to aggregate evidence to identify high-probability interpretations. To test this hypothesis, we fixed the depth of backchaining to 5, and varied the beam and window parameters to investigate their relative importance in finding the global-optimal solution.

To provide a quantitative score of the optimality of the solution, we first identified the exact set of etcetera literals that would need to be assumed to entail both the observations and target inferences in Table 1 using our knowledge base of 81 axioms. Coincidentally, the number of these assumptions was also 81, but some axioms encoding prior probabilities did not participate in the target interpretation, and other axioms were instantiated multiple times. Using this set of 81 target assumptions, we then computed a precision, recall, and F_1 score for a given output with particular parameter settings, considering only the most probable solution of the N-best list. Precision was computed as the number of literals in the solution that can be unified with the target assumptions over the size of the solution. Recall was computed as the number of literals in the target assumptions that could be unified with solution literals, divided by 81.

Table 4 lists the precision and recall scores for beam sizes of 1, 10, 100, and 1000 and window sizes up to 6, reporting all cases where our implementation was able to find a solution in under 600 seconds. No parameter setting was found that could produce the exact target interpretation of the Heider-Simmel film, but these results indicate that larger window sizes are much more instrumental in producing the desired interpretation than larger beams. We surmise that the Hieder-Simmel film is representative of interpretation problems where common factors help explain events that are relatively close together in the input sequence, and that larger window sizes enable these assumptions to explain multiple observations, promoting their relative probability such that they remain on the beam of running interpretations. We find it surprising, however, that exponentially increasing the beam in our experiments had almost no effect on accuracy. We hypothesize that even with a beam size of 1000, the target interpretation is dropped before it can be supported by further evidence, owing to the enormous number of candidate solutions that are considered. We suspect that this hypothesis would be evidenced if it were practical to use beams orders of magnitude larger. However, the completion time of our algorithm scales linearly with the number of contexts considered, limiting the exploration of exponentially larger beams.

8. Conclusions

In the pursuit of humanlike artificial intelligence, automating the processes of narrative interpretation poses several difficult methodological challenges for researchers, owing to the very personal and idiosyncratic nature of sense-making when dealing with narrative content. As the output of these processes are literally “open to interpretation" and largely unobservable, the prospects for curating vast amounts of training data with high levels of inter-rater agreement are slim, and therefore ill-suited to contemporary machine learning methods. On the other hand, knowledge-based approaches to narrative interpretation are also problematic. In addition to new reasoning algorithms, a robust capability for narrative interpretation requires insurmountable amounts of commonsense knowledge. In our research, we mitigate these problems by treating the interpretation as a given, in order to make progress on the reasoning algorithm. It is not at all remarkable that we can devise a knowledge base of hand-authored axioms that is able to replicate an interpretation that we also devised ourselves, produced from an input representation of our own creation. Instead, the contribution of this work is in advancing technologies that allow abductive reasoning to be applied to large interpretation problems.

Our current work focuses on only one movie and one target interpretation, raising questions about the generality of the approach. While the specific knowledge base authored for this research is unlikely to be reused in future applications, we argue that the Incremental Etcetera Abduction algorithm is broadly applicable to interpretation problems consisting of long sequences of input observations. The algorithm employs no domain-dependent heuristics in controlling its search, and allows both the set of observations and the knowledge base to be expressed as literals and definite clauses, respectively, in first-order logic. Interpretations are ordered by their probability, reified as etcetera literals in the antecedents of knowledge base axioms. Although not done in our current research, we envision that these probabilities can be estimated from empirical data in future

applications. Admittedly, our expectations about the generality of this approach must be evidenced by successful applications of this algorithm to new interpretation problems in future research.

One promising area of application is in knowledge-driven natural language understanding, which motivated Hobbs et al.'s (1993) original proposal for interpretation as abduction. Incremental abductive reasoning may help address language understanding problems that remain difficult for contemporary data-driven approaches (e.g., word sense disambiguation), where the intended meanings of individual words are jointly disambiguated by the context of other ambiguous words, the running interpretation of a text, and pragmatic considerations of the discourse. Incremental Etcetera Abduction provides one approach to contextual disambiguation by maintaining running interpretations that are amiable to high-level pragmatic reasoning, all within a framework that offers hope of integration into contemporary probabilistic language processing pipelines.

There are several limitations of the Incremental Etcetera Abduction algorithm that we aim to address in future work. Our current approach of dividing large sequences of observations into equal-sized windows has the benefit of simplicity, but is not well-suited to online, real-time interpretation of incoming observations. Rather than waiting for a window to fill up with incoming observations, a real-time interpretation algorithm should begin consideration of observations as soon as they are made, and do so in a way that avoids redundant computation as subsequent observations arrive. Likewise, a true anytime algorithm for interpretation would scale its search for the most probable interpretation based on the timing of input observations, rather than fixed parameters for beam, window, and depth. When applied in time-critical applications, substantial performance gains could be achieved by rewriting the Python implementation of Incremental Etcetera Abduction in a compiled, strongly-typed language.

As noted in Section 5, our approach can fail to find a more-probable solution due to the skolemization of universal variables in previous windows, preventing unification of literals with a constant in the current window. This limitation reduces the complexity of the software implementation in our current approach, but could be removed with a more clever approach to variable substitution in the running hypotheses on the beam. Lastly, we expect that some applications will require expressivity in the knowledge base beyond what is possible with first-order definite clauses. Adding support for negation, multiple consequents, and argument inequalities create new opportunities to use existing knowledge bases of first-order logical axioms, particularly if these features could be added without severely compromising computational efficiency.

We have found the original Heider-Simmel film to be a rich source of inspiration in the computational modeling of high-level narrative cognition, and see many opportunities for continued investigation of this one short film within AI. One area of future work is in the integration of its high-level interpretation with lower-level computational models of action perception and segmentation, as pursued in Thibadeau's (1986) previous work. In particular, the influence of the film's running interpretations may help explain how there can be high agreement among viewers as to the observed actions (Shor, 1957), despite evidence of low inter-rater agreement when these character movements are viewed out of context (Roemmele et al., 2016). A second area of future work with the Heider-Simmel film concerns the generation of natural language narratives from formal interpretations. While progress has been made in generating fluent English descriptions from similar proof structures (e.g., Ahn et al., 2016), we imagine that the sheer size of this film's interpretation

requires more sophisticated discourse planning methods to produce narratives similar to those of Heider and Simmel's (1944) experimental subjects. Finally, we see the need for future work in the formal representation of commonsense knowledge. Our formalization of the observable events in this film can serve as a starting point for further investigations of its interpretation, to include a richer representation of the emotions, intentions, and relationships that people ascribe to the characters.

Acknowledgements

This research was supported by Grant N00014-16-1-2435 from the Office of Naval Research, which is not responsible for the content and conclusions contained herein.

References

Ahn, E., Morbini, F., & Gordon, A. S. (2016). Improving fluency in narrative text generation with grammatical transformations and probabilistic parsing. *Proceedings of the Ninth International Natural Language Generation Conference* (pp. 70–73). Stroudsburg, PA: Association for Computational Linguistics.

Dennett, D. C. (1989). *The intentional stance*. Cambridge, MA: Bradford Books.

Geib, C. W., & Goldman, R. P. (2009). A probabilistic plan recognition algorithm based on plan tree grammars. *Artificial Intelligence*, *173*, 1101–1132.

Gordon, A. S. (2016). Commonsense interpretation of triangle behavior. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 3719–3725). Palo Alto, CA: AAAI Press.

Greenberg, A., & Strickland, L. (1973). "Apparent Behavior" revisited. *Perceptual and Motor Skills*, *36*, 227–233.

Heider, F. (1958). *The psychology of interpersonal relations*. Mahwah, NJ: Lawrence Erlbaum.

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

Hobbs, J. R., Stickel, M. E., Appelt, D. E., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, *63*, 69–142.

Inoue, N., & Gordon, A. S. (2017). A scalable weighted Max-SAT implementation of propositional etcetera abduction. *Proceedings of the Thirtieth International Conference of the Florida AI Society* (pp. 62–67). Palo Alto, CA: AAAI Press.

International Organization for Standardization (2007). *Common logic (CL): A framework for a family of logic-based languages*. Standard no. ISO/IEC 24707:2007, International Organization for Standardization, Geneva, CH.

Kabanza, F., Filion, J., Benaskeur, A., & Irandoust, H. (2013). Controlling the hypothesis space in probabilistic plan recognition. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 2306–2312). Palo Alta, CA: AAAI Press.

Kautz, H., & Allen, J. F. (1986). Generalized plan recognition. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 32–38). Palo Alto, CA: AAAI Press.

Maslan, N., Roemmele, M., & Gordon, A. S. (2015). One hundred challenge problems for logical formalizations of commonsense psychology. *Proceedings of the Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning* (pp. 107–113). Palo Alto, CA: AAAI Press.

Massad, C., Hubbard, M., & Newtson, D. (1979). Selective perception of events. *Journal of Experimental Social Psychology*, *15*, 513–532.

McCarthy, J. C. (1986). Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, *28*, 89–116.

Meadows, B. L., Langley, P., & Emery, M. J. (2013). Seeing beyond shadows: Incremental abductive reasoning for plan understanding. *Papers from the AAAI 2013 Workshop on Plan, Activity, and Intent Recognition* (pp. 24–31). Palo Alto, CA: AAAI Press.

Mirsky, R., Gal, Y., & Shieber, S. (2017). CRADLE: An online plan recognition algorithm for exploratory domains. *ACM Transactions on Intelligent Systems and Technology*, *8*, 1–22.

Molineaux, M., & Aha, D. W. (2015). Continuous explanation generation in a multi-agent domain. *Proceedings of the Third Annual Conference on Advances in Cognitive Systems* (pp. 1–18). Atlanta, GA: Cognitive Systems Foundation.

Roemmele, M., Morgens, S.-M., Gordon, A. S., & Morency, L.-P. (2016). Recognizing human actions in the motion trajectories of shapes. *Proceedings of the Twenty-First International Conference on Intelligent User Interfaces* (pp. 271–281). New York: Association for Computing Machinery.

Shor, R. (1957). Effect of pre-information upon human characteristics attributed to animated geometric figures. *Journal of Abnormal and Social Psychology*, *54*, 124–126.

Thibadeau, R. H. (1986). Artificial perception of actions. *Cognitive Science*, *10*, 117–149.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (pp. 1556–1559). Paris, France: European Language Resources Assocation.