A Appendix

A.1 Event Extraction

We balance the number of content words to ensure that the events are generalizable but still concrete enough to be labelled. We only keep events with at least two and less than five content words, defined as words that are not stop words, person tags, or blanks. We count phrasal verbs (such as "get up") as content word. We limit the sets of events to those events that occur most frequently in our corpora, using corpus-specific thresholds. [5]

A.2 Annotation Setup

Each event was presented to three different raters recruited via Amazon Mechanical Turk. Raters were given the option to say that the event did not make sense (invalid), at which point they were not asked any other questions. If the rater marked the event as valid, they were required to answer the question about how PersonX typically feels after the event. Each rater was paid $0.10 per event. Additionally we annotated a small number of events where "It" was in the subject (e.g., `It rains all day`). For these events, we only asked raters to say how other people typically feel after the event (if they marked the event as valid).

B Event2Mind Training Details

In our experiments, we use Adam to train for ten epochs, as implemented in Tensorflow (Abadi et al., 2015).

For baseline models, the dimension of the event encoded embedding is $H = 300$. For our BiRNN model, we also experimented with an embedding dimension of $H = 100$.

We define the vocabulary as the tokens appearing in the training data events and annotations at least twice, plus the bigrams and trigrams that appear more than five times. In cases where an annotation for the intent/reaction was left blank (because there was no intent or the event did not affect other people), we treated the annotation as equivalent to the word "none". Because many of the annotations for intent started with "to" or "to be", we stripped these two words from the beginning of all intent annotations.

[5] For ROC Story and Spinn3r events, we choose events with frequency at least five and 100, respectively. For Syntactic Ngrams, we took the top 10000 events.

Overlap criterion	% of Event2Mind events
Any node	25%
All annotations, with select relations	12 %
XIntent, with select relations	3%
XReact/OReact, with select relations	<1%

Table 5: Event2Mind events overlap with ConceptNet events. While a non-trivial amount are represented in some capacity, few events have intent or reactions.

C Comparison with ConceptNet

We match our events with the event nodes in ConceptNet and find 6 ConceptNet relations that compare to our intent and reaction dimensions. Specifically, we compare *MotivatedByGoal, CausesDesire, HasFirstSubevent,* and *HasSubevent* with the 'XIntent' annotations, and 'XReact' and 'OReact' annotations with the *Causes* and *HasLastSubevent* relations. For each ConceptNet event, we then compute unigram overlap between our annotations and their ConceptNet proxy using the 6 relations.

We summarize overlap in Table 5, where we show that 75% of Event2Mind events are not covered in ConceptNet. We also show that while 12% of our events have an edge with one of the 6 relations, the actual overlap between our annotations and the ConceptNet data is very low (<5%). This overlap statistics indicates that our dataset provides new commonsense knowledge that is not covered by previous resources such as ConceptNet.

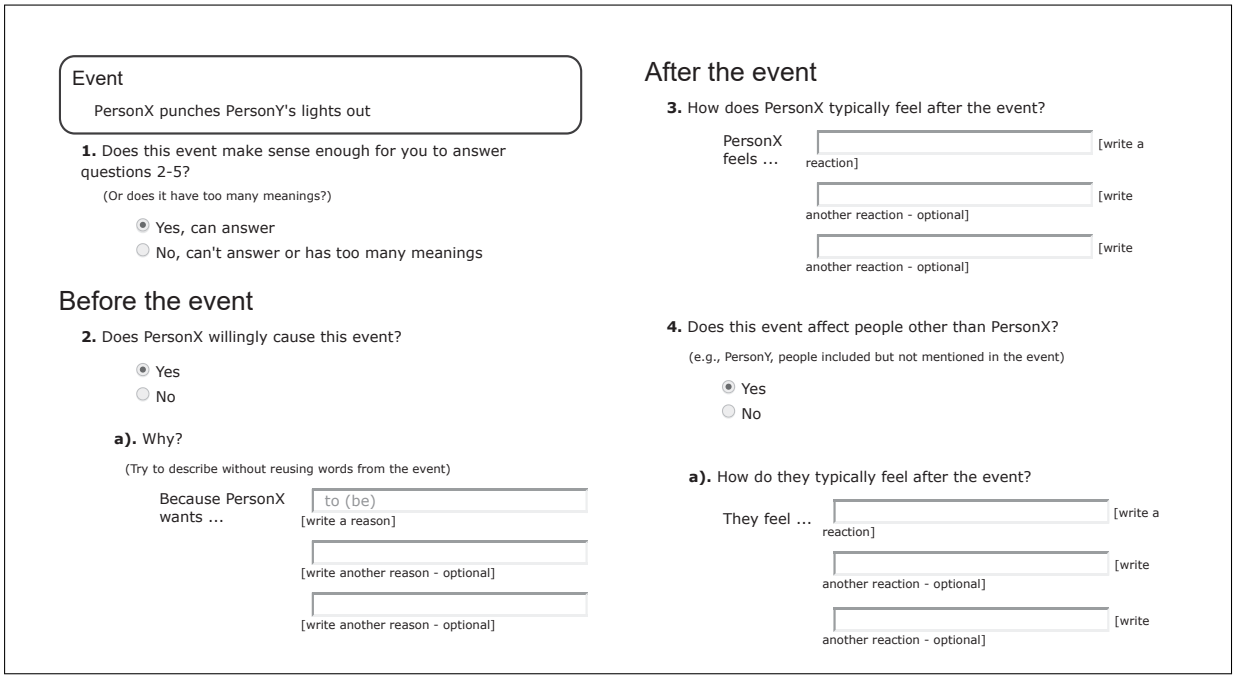

Figure 8: Main event phrase annotation setup. Each event was annotated by three Amazon Mechanical Turk raters.