Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing

Alon Hafri[1], John C. Trueswell[1], and Brent Strickland[2]

[1] Department of Psychology, University of Pennsylvania, 425 S. University Avenue
Philadelphia, PA 19104 USA

[2] Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University
Institut Jean Nicod, (ENS, EHESS, CNRS)
75005 Paris, France

Running Head : Spontaneous encoding of event roles from visual scenes

Correspondence : Alon Hafri
Department of Psychology
University of Pennsylvania
425 S. University Avenue
Philadelphia, PA 19104

E-mail : ahafri@sas.upenn.edu

Word Count : 13480 [175 (Abstract), 1355 (Introduction), 7655 (Methods and Results), 3394 (Discussion), 733 (Footnotes), 32 (Acknowledgements), 136 (Appendix)]

Version : Accepted for publication in *Cognition* on 2/5/2018

Abstract
A crucial component of event recognition is understanding event roles, i.e. who acted on whom: *boy hitting girl* is different from *girl hitting boy*. We often categorize Agents (i.e. the actor) and Patients (i.e. the one acted upon) from visual input, but do we rapidly and spontaneously encode such roles even when our attention is otherwise occupied? In three experiments, participants observed a continuous sequence of two-person scenes and had to search for a target actor in each (the male/female or red/blue-shirted actor) by indicating with a button press whether the target appeared on the left or the right. Critically, although role was orthogonal to gender and shirt color, and was never explicitly mentioned, participants responded more slowly when the target's role switched from trial to trial (e.g., the male went from being the Patient to the Agent). In a final experiment, we demonstrated that this effect cannot be fully explained by differences in posture associated with Agents and Patients. Our results suggest that extraction of event structure from visual scenes is rapid and spontaneous.

Keywords
thematic roles; argument structure; scene perception; event perception; action perception; core cognition

1. Introduction

In order to successfully navigate a perceptually chaotic world and share our understanding of it with others, we must not only extract the identity of people and objects, but also the roles that they play in events: *Boy-hitting-girl* is very different from *girl-hitting-boy* even though the event category (i.e. *hitting*) and actors involved are the same. In the former, the boy is the Agent (the actor) and the girl the Patient (the one acted upon), while in the latter, their roles are reversed. The fundamental importance of such "thematic roles" has long been emphasized in linguistics: Theories of thematic roles were initially developed to account for the consistent semantic properties of grammatical arguments (e.g., *Subjects* and *Objects*) across linguistic descriptions of events (Croft, 2012; Dowty, 1991; Fillmore, 1968; Gruber, 1965; Kako, 2006; Levin & Rappaport-Hovav, 2005) but now they are also a component of some theories of conceptual representation (Jackendoff, 1990; Langacker, 1987; Talmy, 2000), development (Baillargeon et al., 2012; Leslie, 1995; Muentener & Carey, 2010; Yin & Csibra, 2015), and perception (Leslie & Keeble, 1987; Strickland, 2016) more generally.

1.1. Event role extraction

While there is ongoing debate within linguistics about the precise number and nature of thematic roles in language, here we are interested in whether the mind, independently from explicit language production and comprehension tasks, rapidly and spontaneously extracts role information from perceptual input. Our work takes inspiration from a wealth of previous literature that has demonstrated rapid and bottom-up encoding of semantic content from visual scenes. These studies have revealed that categories of both objects (Biederman, Blickle, Teitelbaum, & Klatsky, 1988; Biederman, Mezzanotte, & Rabinowitz, 1982; Thorpe, Fize, & Marlot, 1996) and places (Oliva & Torralba, 2001; Potter, 1976) can be recognized from brief displays (sometimes as little as 13 ms); that the computation itself is rapid – occurring within 100-200 ms (VanRullen & Thorpe, 2001); and that the computation is relatively automatic (Greene & Fei-Fei, 2014).

In previous work we have shown that, just as with object and place categories, event category and event role information is in principle available in a bottom-up fashion from very brief displays (Hafri, Papafragou, & Trueswell, 2013; see also Dobel, Diesendruck, & Bölte, 2007; Glanemann, Zwitserlood, Bölte, & Dobel, 2016; Wilson, Papafragou, Bunger, & Trueswell, 2011). However, it is not yet known whether encoding of event information is rapid: all tasks in previous studies (to our knowledge) explicitly required participants to make a post-stimulus judgment about what was happening in the scene. Thus, the computation itself (although based on a briefly presented visual stimulus) could conceivably have continued for several seconds, up until response to the post-stimulus probe. Additionally, the computation might have occurred only because of the explicit demands of the task, rather than being spontaneous.

1.2. Spontaneity and generality of role encoding

Here, we define a spontaneous process as any process that is executed independently of an explicit goal. Such a process could be automatic, in the sense that it is mandatory given certain input characteristics (Fodor, 1983), but it could also be spontaneous but not automatic in the sense that, under some conditions and with some cognitive effort, the process could be prevented from being executed (Shiffrin & Schneider, 1977). In the present work, we test for spontaneity of event role encoding.

Given the particular importance of event roles to event understanding, the spontaneity of such a process would be beneficial as we engage the social world, since at any given moment we may be performing other perceptual tasks, e.g., identifying objects or spatial properties of the scene. It would also prove useful to the young language learner tasked with mapping utterances to the events that they refer to (a problem discussed in detail in Gleitman, 1990; Pinker, 1989).

In both of these situations (social and linguistic), the utility of role information arises from its relative generality, i.e., the identification of commonality between the actors engaged in different events, such as *kicking* and *pushing* (with the degree of commonality perhaps dependent on abstract properties shared between the participants in these events, such as volition or cause; Dowty, 1991; Jackendoff, 1990; Pinker, 1989; Talmy, 2000; White, Reisinger, Rudinger, Rawlins, & Durme, 2017). However, research on action recognition using psychophysical and neuroscientific methods has largely focused on how the perceptual system differentiates between different action categories (e.g., *kicking, pushing, opening*) and generalizes within action category (Hafri, Trueswell, & Epstein, 2017; Jastorff, Begliomini, Fabbri-Destro, Rizzolatti, & Orban, 2010; Oosterhof, Tipper, & Downing, 2012; Tucciarelli, Turella, Oosterhof, Weisz, & Lingnau, 2015; Wurm & Lingnau, 2015). This research has not yet addressed how we come to recognize the distinct roles that multiple actors play in visual scenes, or how (and whether) our perceptual system generalizes across the agents of different actions.

Investigating the perception of events in visual scenes provides an ideal avenue to test hypotheses about the generality of event roles. One hypothesis is that awareness of event-general properties of event roles (e.g., volition or cause) arise through explicit and deliberate observation of commonalities among event-specific roles (e.g., *kicker, kickee*) outside of the domain of perception (Tomasello, 2000). However, to the degree that we can find evidence that perception itself rapidly and spontaneously furnishes such event-general role information, the notion of event-specific roles as drivers of event understanding from scenes becomes less plausible. We hypothesize that in initial scene viewing, the perceptual system rapidly categorizes event participants into two broad categories – "Agent-like" and "Patient-like" (denoted Agent and Patient from here on for simplicity; Dowty, 1991; Strickland, 2016) – even if these assignments are later revised or refined in continued perceptual or cognitive processing of the event (see section 6.1 for elaboration on these issues).

1.3. The current study: an event role switch cost?

The goal of the current work is to establish the degree to which the visual system gives the observer event roles "for free", as part of routine observation of the world. We aim to show the following: (1) that the visual system encodes event roles spontaneously from visual input, even when attention is otherwise occupied (i.e. even when the observer is not explicitly asked to recognize events but rather is engaged in some orthogonal task); (2) that the computation of role

itself is rapid; (3) that this encoding of event roles is at least partly event-general; and (4) that any evidence we find for encoding of event roles cannot be fully accounted for by simple visual correlates of event roles alone, such as posture.

To achieve this goal, we employed a "switch cost" paradigm (Oosterwijk et al., 2012; Pecher, Zeelenberg, & Barsalou, 2003; Spence, Nicholls, & Driver, 2001). In several experiments, participants observed a continuous sequence of two-person scenes and had to rapidly identify the side of a target actor in each (Experiments 1a and 1b: male or female actor; Experiments 2 and 3: blue- or red-shirted actor). With our design, event role identities provide no meaningful information for the primary task of gender or color identification, so observers need not attend to such irrelevant information. Nevertheless, we hypothesized that when people attend to the target actor to plan a response, then if event roles are spontaneously encoded, they should "come along for the ride." Thus, we should be able to observe an influence of this role encoding on responses even though event roles are irrelevant to the primary task.

More specifically, we reasoned that if role assignment is spontaneously engaged, then when the role of the target actor switched from trial to trial, it would result in a cost, i.e., a relative lag in reaction time, even though subjects were tasked with identifying a property orthogonal to roles (here, gender or shirt color). If such a pattern were observed, it would provide compelling evidence that analysis of event structure from visual scenes is a rapid, spontaneous process that is engaged even when we are attending to other perceptual information. Furthermore, by using simple tasks based on visual information known to be rapidly available (including gender; Mouchetant-Rostaing, Giard, Bentin, Aguera, & Pernier, 2000), we expected that observers would respond quickly, allowing us to test the rapidity of extraction of event role information.

2. Experiment 1a

Participants observed a series of simple still images displaying an interaction between a male and a female, and were simply asked to say whether the male/female was on the left or right of the screen. We predicted that although the task fails to actively encourage role encoding (and may even discourage it), participants would nevertheless be slower on trials in which the event role of the target actor differed from his or her role in the previous trial, i.e., a "role switch cost".[1]

2.1. Method

2.1.1. Participants

Twenty-four members of the University of Pennsylvania community participated and received

[1] We cannot differentiate between switch costs vs. repetition benefits (priming) because there is no baseline for comparison, but in keeping with the terminology in previous investigations using this paradigm (e.g., Pecher et al., 2003), we use the term switch costs. Whether the effects are a benefit or cost does not qualitatively change our conclusions.

either class credit or $10. Because we were collecting a large number of trials within-participant (see section 2.1.3 below), we predicted that this number of participants would be sufficient to observe the role switch cost, if it were to exist. All participants in this experiment and in the other experiments reported below gave informed consent, following procedures approved by the University's institutional review board.

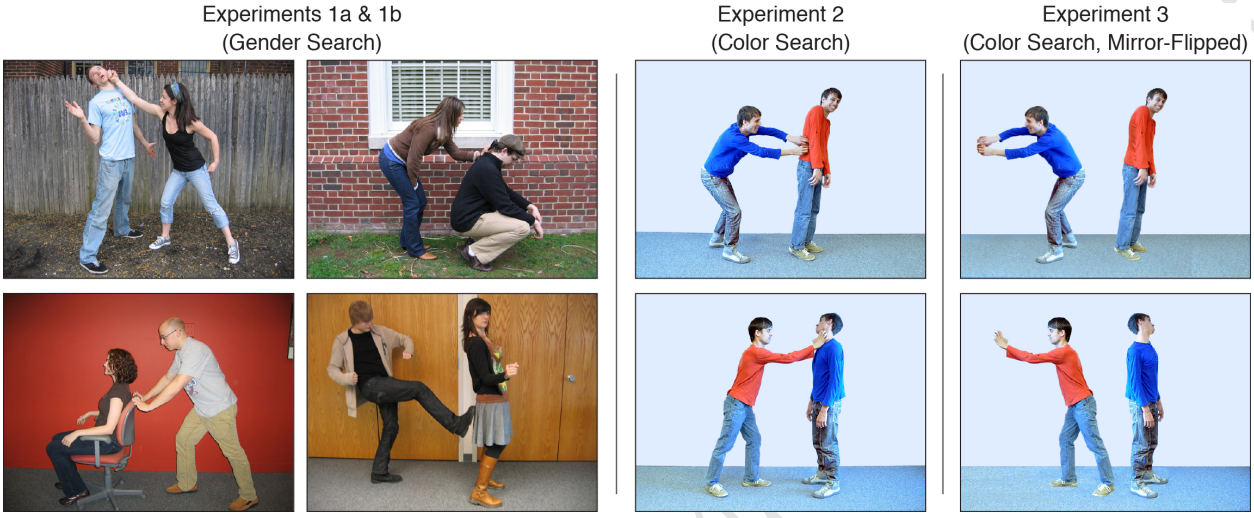


Fig. 1. Example stimuli. All experiments featured 10 event categories. In Experiments 1a and 1b, these were depicted by several different pairs of actors, and Agent gender (male or female) and Agent side (left or right) were fully crossed within event category. In Experiments 2 and 3, events were depicted by a pair of identical twin actors. Agent shirt color (blue or red) and Agent side (left or right) were fully crossed within event category. In Experiment 3, the images from Experiment 2 were manipulated such that the two actors were always facing opposite one another; thus, their interactive relationship was almost entirely eliminated. See the Appendix for more example images.

2.1.2. Materials

The stimuli were 40 color photographic images depicting 10 two-participant event categories taken from a previous study that investigated extraction of event categories and roles from briefly displayed and masked images (Hafri et al., 2013). The event categories used were *brushing, chasing, feeding, filming, kicking, looking, punching, pushing, scratching, tapping*. These categories were chosen because they showed the highest agreement among subjects for role assignment from brief display (i.e., male as Agent or Patient). All stimuli were normed for event category and role agreement in the previous study.

Six different male-female actor pairs appeared in the images, with each actor pair appearing in front of a different indoor or outdoor scene background. Each event category was associated with only one of the actor pairs (e.g., *brushing* and *chasing* was always performed by Pair 1, *feeding* by Pair 2, etc.). For each event category, the gender of the Agent (male or female) and the side of the Agent (left or right) were fully crossed, such that there were four stimuli for each event category. Each event depicted the actors in profile view. Example images appear in Figure 1, and examples for each event category appear in the Appendix.

For all experiments, images were 640 × 480 pixels and subtended 19° × 15° at approximately

54 cm distance from the screen. Stimuli were displayed on a 19" Dell 1908FP LCD monitor at a refresh rate of 60 Hz. Responses were collected using a PST E-Prime button box (mean latency 17.2 ms, *SD* 0.92 ms, measured in-lab). The experiment was run in Matlab using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

2.1.3. List design

Given that detecting switch costs depends on measuring the influence of one stimulus on another, we implemented "continuous carryover" sequences, which are similar to randomized block and Latin square designs, with the added benefit of controlling for first-order carryover effects, i.e. each stimulus precedes and follows every other stimulus (Aguirre, 2007; Nonyane & Theobald, 2007). This design resulted in 1601 trials split among 40 blocks. Unique lists were generated for every participant. An additional reason we used this list design was that it naturally provided a large number of trials with which to precisely measure effects of all factors manipulated in the experiment, across both subjects and items. This was important: given that participants were actively required to attend to stimulus features orthogonal to the property of interest (event roles), there was potential for the role switch cost to be quite subtle.

To maximize our chances of finding a switch cost if it were to exist, a small number of catch trials (Event Test trials) were randomly dispersed among the standard image trials. On these catch trials, participants were given a 2AFC test on what action just appeared in the previous trial (e.g., *kicking* or *pushing*). One label was correct, and the other was a foil randomly selected from the set of nine other categories. There were 58 catch trials in total, with 1 to 3 per 40-trial block.

2.1.4. Procedure

Subjects were instructed that as each image appeared, they would have to press one of two buttons (left or right) to indicate, as quickly and accurately as possible, which side of the screen that the target actor appeared on (left button for left, right button for right). For half of the subjects, the target was the male actor, and for the other half, the female actor (i.e. male or female search was between-subject, counterbalanced across participants). There were 40 blocks of trials, each of which was a continuous sequence of all 40 image trials and the interspersed catch trials, followed by a quick break before the next block. The purpose of the catch trials was to focus participants' attention on the events they were observing without explicitly testing them on event roles (see section 2.1.3 above). Subjects were told that they would be intermittently tested on what action just appeared in the previous trial.

Figure 2 illustrates the trial and block sequence. Each trial consisted of the following: A "Ready?" screen for 350 ms, a central fixation crosshair for 250 ms, a blank screen for 150 ms, and the test image, which remained on the screen until the subject responded. Catch trials involved a similar sequence, but in place of the test image was a slide with the text "What action did you just see?" and two event category labels on either side of the screen (e.g., "biting" and "pushing"). Subjects selected their answer by pressing either the left or right button. Image trials timed out if no response was given within 2000 ms, and catch trials within 3500 ms. Twelve

practice image trials and two catch trials preceded the main experimental sequence. Practice image trials depicted joint or symmetrical actions (e.g., *crying, shaking hands*). Average duration of the experiment was 41 min (which was similar across all additional experiments reported below).

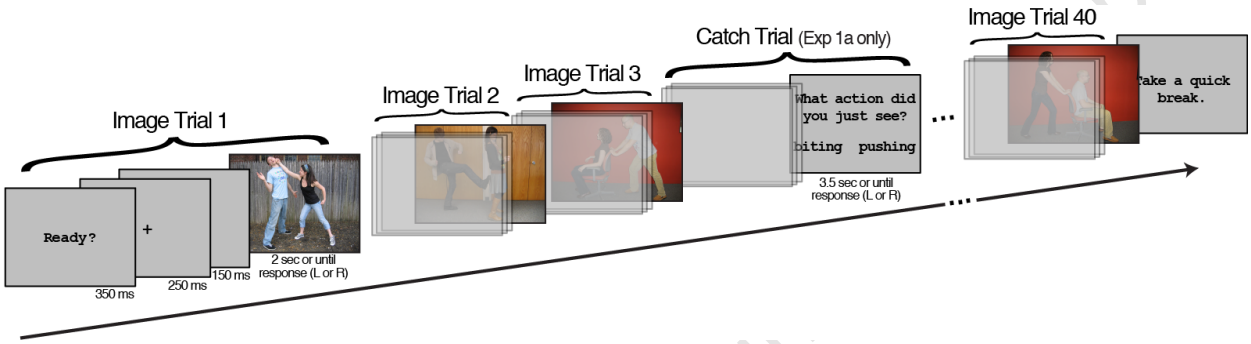


Fig. 2. Block structure for all experiments. On each image trial, subjects pressed a button to indicate the position of the target actor as fast as possible (left or right). In Experiments 1a and 1b, the target actor was the male or female (between-subject). In Experiments 2 and 3, the target actor was the blue- or red-shirted actor (between-subject). Only Experiment 1a contained catch trials, which asked about the action that appeared in the previous trial.

2.1.5. Data analysis

Trial exclusion criteria were decided in advance of analysis and were the following: trials with an incorrect response and those following an incorrect trial, RTs faster than 200 ms, timeouts, trials after breaks, and trials after catch trials. An additional 63 trials in total across all subjects were also excluded due to an error in list creation. For the remaining data, trials with RTs 2.5 standard deviations above or below each subject's mean were also excluded, following accepted data trimming procedures (e.g., Balota, Aschenbrenner, & Yap, 2013). A mean of 17% (*SD* 4.0%) of trials in total were excluded per subject, which meant there were an average of 269 trials included per subject. Average accuracy was 95.6% (*SD* 2.2%), and average RT on image trials for the included data was 383 ms (*SD* 34 ms).

Individual trial reaction times from the primary task (i.e., judging gender side) were analyzed with linear mixed effects modeling using the lme4 R package (Bates et al., 2016), with centered (sum-coded) predictors. The analyses used the maximal subject and item random effects structure that converged for all tested models (Barr, Levy, Scheepers, & Tily, 2013).[2] RTs were

[2] When more complex random effects structures failed to converge, we successively dropped random slope terms with the smallest variance, until the model converged (Barr et al., 2013). The random effects structures used for each experiment and cross-experiment comparison were the following (in R model syntax):

- Experiment 1a: (1+Actors+Side|subjNum)+(1+propertyAgent*sideAgent|event)
- Experiment 1b: (1+Actors*Role+Actors*Side|subjNum)+(1+propertyAgent*sideAgent| eventCategory)

first transformed into inverse RTs (-1000/RT) to improve normality for model fitting. Additionally, all models included nuisance regressors for trial number and preceding trial inverse RT to account for general temporal dependencies (Baayen & Milin, 2010).

The following factors were included in models: Actors (repeated vs. switched), i.e., whether the actor pair was the same or different from the previous trial; Side (repeated vs. switched), i.e., whether the side of the target actor (e.g., male) was the same or different as the previous trial; and the effect of primary interest, Role (repeated vs. switched), i.e., whether the role of the target actor was the same or different (e.g., whether the male remained the Agent or switched to being Patient). Significance of these factors was tested by comparing likelihood-ratio values for nested models that included main effects and interactions of factors to models without them.[3]

2.2. Results

2.2.1. Role switch cost

An event role switch cost was observed. As shown in Table 1, participants were on average 6 ms slower when the role of the target character changed from one trial to the next. This effect, though quite small, was significant: The best-fitting mixed effects model included a main effect of Role (the role switch cost) and main effects and interactions of Actors and Side. The fit of this model was significantly better than a model without the main effect of Role, $\chi^2(1) = 52.9$, $p < .001$. Models with additional interaction terms were not a significantly better fit, either for Actors × Role ($\chi^2(1) = 1.71$, $p = .19$), or Side × Role ($\chi^2(1) = 0.09$, $p = .76$). See Table 1 for a summary of the effects from the best-fitting model.

2.2.2. Absolute vs. relative magnitude of role switch cost

Before continuing, we believe that the empirical robustness and theoretical import of the role switch cost must be separated from the absolute size of the effect observed. Although the

- Comparison of Experiments 1a and 1b: (1+Role|subjNum)+(1+propertyAgent*sideAgent| eventCategory)
- Experiment 2: (1+Role*Side|subjNum)+ (1+propertyAgent*sideAgent|eventCategory)
- Experiment 3: (1+Role|subjNum)+(1+propertyAgent*sideAgent|event)
- Comparison of Experiments 2 and 3: (1+Role|subjNum)+(1+propertyAgent*sideAgent| eventCategory)

Abbreviations (consistent for all experiments): subjNum = subject identity; propertyAgent = Agent gender (Male or Female, Experiments 1a and 1b only), or Agent Color (Blue or Red, Experiments 2 and 3 only); sideAgent = Agent side (Left or Right); eventCategory = event category (e.g., *kicking*).

[3] Here and in Experiment 1b, repeated event always entailed repeated actors, due to the nature of the stimuli employed (see section 2.1.2). However, similar results were obtained with Event as a factor instead of Actors. Likewise, since actors and scene backgrounds co-varied, Actor switch entails a Background switch (and vice-versa), but for simplicity, we will refer to this factor as same/different Actors.

absolute magnitude of the role switch cost was small (about 6 ms), the *standardized* effect sizes were quite large: Cohen's *d* of 1.07 and 2.24, for subjects and items, respectively (see Figure 3). As another indication of its robustness, 21/24 participants and all 10 event categories showed a numerical difference in line with the role switch cost. And while it may be surprising that such a small effect would be statistically significant, each observer provided on average 1329 data points, resulting in very stable performance estimates per subject and per item (e.g., note the tight 95% confidence intervals across subjects in Table 1). Furthermore, it is within the same order of magnitude of previously observed switch costs, relative to mean RTs for task: for example, Pecher et al. (2003) obtained a cost of 29 ms relative to mean RTs of 1139 ms (a ratio of 2.5%), and Oosterwijk et al. (2012) obtained a cost of 22 ms relative to mean RTs of 1683 (a ratio of 1.3%), compared with our 6 ms vs. 383 ms mean RTs (a ratio of 1.6%). Similar arguments hold regarding the absolute vs. relative magnitude of the role switch cost observed in the other experiments reported in this manuscript, and we return to this issue in section 6.6.

2.2.3. Other observed effects

Besides the effect of primary interest (event roles), the best fitting model revealed several additional effects. First, people were slower when Actors switched. This is not surprising: when actor pair switched, it likely took longer to ascertain which character was the male or female. There was also an interaction of Side × Actors: On trials where the actor pair was different, participants were faster when the target side switched. Though speculative, it may be that with a significant visual change such as a switch in the actors, subjects may have expected a side switch, resulting in a switch benefit, or faster RTs. Whatever the reason for these additional effects, the role switch cost was invariant to these other factors (Side and Actors).

2.2.4. Event catch task

Average RT on catch trials was 1177 ms (*SD* 215 ms), and accuracy on catch trials was significantly above chance across participants (mean = 85%, *SD* = 10%, $t(23) = 40.0$, $p < .001$, $d = 3.37$, 95% CI = [81%, 89%]). This indicates that participants were monitoring the events in the images sufficiently to distinguish which of two event categories they observed in the previous trial.

One important question is whether event category extraction is related to event role extraction. Although in our previous work we found that role recognition was not significantly correlated with event category extraction on an item-by-item basis (Hafri et al., 2013), we can also address this in the current study, in two ways. First, if there is a relationship between event category and event role extraction, we might find that the magnitude of the role switch cost is correlated on a subject-by-subject basis with performance on catch trials (event identification). However, we found no significant correlation between individual participants' role switch cost magnitude (based on the mean inverse RT difference between repeated and switch role trials for each subject), and either their overall accuracy on catch trials, $r = -0.11$, $t(22) = -0.52$, $p = .61$, or their mean inverse RT on catch trials (accurate trials only), $r = 0.00$, $t(22) = -0.01$, $p = .99$.

Another way to investigate the relationship between event category and event role extraction is

by asking whether catch trial (event identification) performance would be worse when the catch trial probe is about an image in which event role switched. To assess this, we split catch trials by whether the previous trial was a Repeated or Switched Role image trial (an average of 27.8 trials in each condition per subject, range 20-36). We ran multilevel models to predict performance (either accuracy or inverse RT) on catch trials across subjects. Specifically, we tested whether adding a main effect of Previous Role (Repeated vs. Switched) to the models would improve model fit, over a null model without the Previous Role main effect (both models included a random intercept and random slope for Previous Role for each subject). However, adding Previous Role did not significantly improve fit either for catch trial accuracy (logistic regression, $\chi^2(1) = 0.64$, $p = .42$) or for catch trial inverse RT ($\chi^2(1) = 3.09$, $p = .08$); and even though improvement for the inverse RT model was marginal, it went in the opposite direction of the prediction, i.e. faster RTs on catch trials when the previous trial role *switched*.

Although these tests are post-hoc and we should interpret the null results with caution, they at least imply that at the individual subject or trial level, event category identification is robust to changes in role information. Nevertheless, a more definitive test of the relationship between event role and category extraction would require further experimentation.

Table 1: Mean RTs across subjects for Experiment 1a, separately for all factors that were significant in model fitting (significant interaction terms split by each factor level). 95% confidence intervals in parentheses.

Condition	Reaction time (ms)		Switch cost (ms)	*t* value for parameter in best-fitting model
	Repeated	Different		
Role	380 (14.2)	386 (14.8)	6 (2.00)	7.27*
Actors	371 (12.8)	385 (14.9)	14 (3.68)	3.99*
Side	390 (16.0)	377 (13.7)	-13 (6.26)	-3.95*
Side, Repeated Actors	371 (12.9)	371 (13.5)	0 (6.39)	0.47
Side, Switched Actors	393 (16.5)	378 (13.9)	-15 (6.55)	-3.95*

* $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 2.2.1 for details on model comparisons.

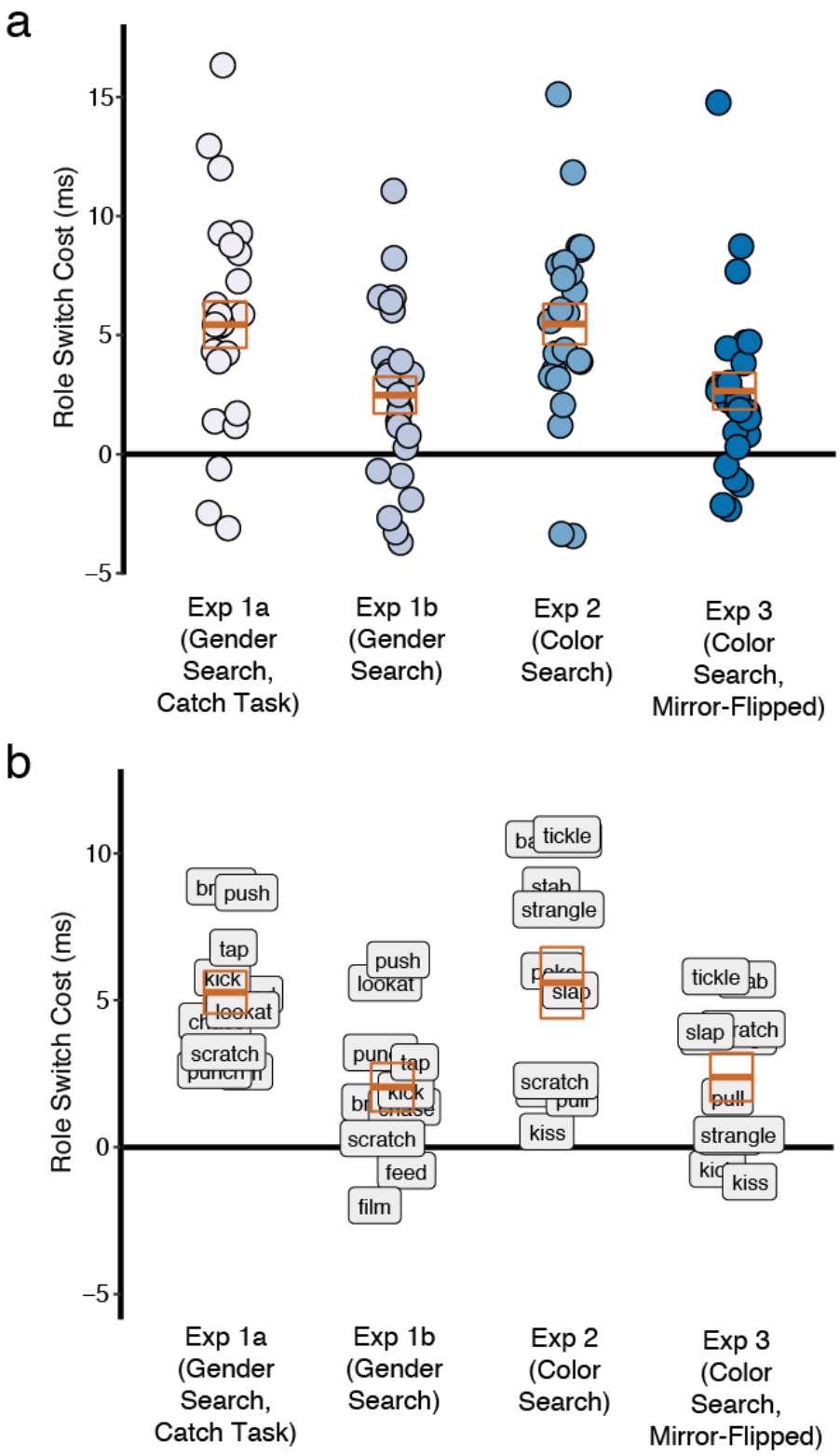


Fig. 3. Individual (a) subject and (b) item (event category) means for the role switch cost, across all experiments. These plots show the consistency of the role switch cost for both subjects and items: the majority of means are above zero in each case. Orange boxes indicate the mean and standard error across subjects and items, for each experiment.

2.3. Discussion

Although a role switch cost was observed in Experiment 1a, the Event Test catch trials may have inadvertently focused attention on event roles. Experiment 1b was identical to the previous experiment, except that there was no catch task and no mention of events or actions. If this effect is really a result of the default in visual perception of scenes, then we expected to observe it even under these conditions.[4]

3. Experiment 1b

The Event Test catch trials in Experiment 1a may have inadvertently focused attention on event roles. The current experiment was identical to Experiment 1a, except that there was no catch task and no mention of events or actions.

3.1. Methods

3.1.1. Participants

An additional 24 members from the University of Pennsylvania community participated and received class credit. Given the stability of the role switch effect in Experiment 1a, we believed this number to be sufficient.

3.1.2. Materials and procedure

All materials, apparatus, and procedure were identical to Experiment 1a, except that no catch (Event Test) trials were included, and instructions were modified to omit mention of the catch task or actions and events.

3.1.3. Data analysis

Data coding, trial exclusion criteria, and analysis were the same as in Experiment 1a. An additional 216 trials across all subjects were excluded due to an error in list creation. A mean of 13% (*SD* 4.9%) of trials (214 on average) per subject were excluded, average accuracy was 96.0% (*SD* 2.6%), and average RT for the included data was 387 ms (*SD* 48 ms). Individual trial RTs

[4] Preliminary analyses of Experiments 1a and 1b originally appeared in conference proceedings (Hafri, Trueswell, & Strickland, 2016).

were analyzed using linear mixed effects modeling.

3.2. Results

As in Experiment 1a, a role switch cost was observed. In Table 2, we see that participants were on average 3 ms slower when the role of the target character changed from one trial to the next. This effect was once again robust: 17/24 subjects and 7/10 event categories went in the direction of the effect (Cohen's *d* of 0.55 and 0.58, respectively; see Figure 3). And although small, it was significant: The best-fitting mixed effects model included main effects and interactions of Role and Actors, as well as a main effect of Side and the interaction of Side × Actors. The fit of the model was significantly better than the same model without the additional interaction of Role × Actors, $\chi^2(1) = 4.89$, $p = .03$, and significantly better than a model that did not include Role at all, $\chi^2(2) = 15.5$, $p < .001$. A model with an additional interaction of Role × Side was not a significantly better fit, $\chi^2(1) = .004$, $p = .95$.

Interestingly, the role switch cost was greater when the actor pair repeated than when it did not, although importantly, the role switch cost was significant even when the actor pair differed. And as in Experiment 1a, on trials where the actor pair was different, participants were *slower* when the side repeated. See Table 2 for details.

Table 2: Mean RTs across subjects for Experiment 1b, separately for all factors that were significant in model fitting (significant interaction terms split by each factor level). 95% confidence intervals in parentheses.

Condition	Reaction time (ms)		Switch cost (ms)	*t* value for parameter in best-fitting model
	Repeated	Different		
Role	385 (19.9)	388 (20.3)	3 (1.59)	2.62*
Actors	371 (16.8)	390 (20.8)	19 (5.55)	2.08*
Side	394 (19.5)	380 (21.2)	-14 (7.55)	-5.09*
Role, Repeated Actors	368 (16.8)	374 (17.2)	6 (5.65)	3.60*
Role, Switched Actors	388 (20.5)	391 (21.0)	3 (1.84)	2.62*
Side, Repeated Actors	368 (14.0)	374 (20.2)	6 (11.6)	0.92
Side, Switched Actors	398 (20.6)	382 (21.5)	-16 (7.26)	-5.09*

* $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 3.2 for details on model comparisons.

3.2.1. Comparison of Experiments 1a and 1b

Not surprisingly, more participants and items showed the numerical difference in Experiment 1a (with the catch task) than in Experiment 1b (without the catch task; 21/24 vs. 17/24 participants, and 10/10 vs. 7/10 items, respectively; see Figure 3). To formally compare the two experiments, we ran new mixed effects models with the data from both experiments combined, starting with a base model whose main effects and interactions were identical to the best-fitting

model of Experiment 1b. The best-fitting model in this combined analysis had main effects of Actors, Side, Role, and Experiment, and interactions of Actors × Side, Role × Actors, Role × Experiment, and Actors × Experiment. The fit of the model was significantly better than a model without the additional interaction of Role × Experiment, $\chi^2(1) = 3.88$, $p = .05$. The greater role switch cost for repeated actors vs. switched actors observed in Experiment 1b appears to be consistent across both Experiments 1 and 1b: adding the triple interaction of Role × Actors × Experiment to the best-fitting model in the current analysis did not significantly improve the fit, $\chi^2(1) = 0.74$, $p = .39$. This analysis confirms that the role switch cost was indeed greater in Experiment 1a than in Experiment 1b.

Additionally, items drove the role switch cost consistently across experiments: the role switch costs for individual image stimuli were correlated across experiment, $r = 0.37$, $t(38) = 2.43$, $p = .02$. This correlation further attests to the stability of the measures of central tendency (i.e., subject and item means) – likely due to the large number of observations per image.

4. Experiment 2

In this experiment, we tested the generalizability of the role switch cost. We ran the same paradigm of Experiment 1b, with two changes: (1) we used new event categories and stimuli, in which events were depicted by a pair of red- and blue-shirted identical twin actors; and (2) the main task was to identify the side of the blue or red-shirted actor. As in Experiment 1b, there was no catch task and no mention of events or actions. If spontaneous role assignment is really the default in scene perception, then we expected to observe the role switch cost even with these changes.

4.1. Methods

4.1.1. Participants

An additional 24 members from the University of Pennsylvania community participated and received class credit. Given the stability of the role switch effect in Experiments 1a and 1b, we believed this number to be sufficient. Data from an additional three participants were excluded: one for a high number of sub-200 ms RTs (145 trials), one for non-completion, and one for falling asleep.

4.1.2. Materials

Stimuli were 40 color photographic images depicting 10 two-participant event categories, taken from a previous study (Hafri et al., 2013): *bandaging, kicking, kissing, poking, pulling, scratching, slapping, stabbing, strangling, tickling*. All categories except *kicking* and *scratching* differed from those used in Experiments 1a and 1b, providing a test of the

generalizability of the role switch cost to new event categories. All stimuli were normed for event category and role agreement in the previous study, and showed high agreement for event role extraction from brief display. Events were depicted by a single pair of identical-twin actors (male, age 29) who dressed the same except for a difference in shirt color (blue vs. red). As in Experiments 1a and 1b, for each event category, the shirt color of the Agent (blue or red) and the side of the Agent (left or right) were fully crossed, such that there were four stimuli for each category. Example images appear in Figure 1, and examples for each event category appear in the Appendix.

4.1.3. Procedure

Apparatus, list design, and procedure were identical to Experiment 1b, except that that the words "male" and "female" were replaced by "blue" and "red" in the instructions.[5] Task (blue or red search) was between-subject, counterbalanced across participants. Sixteen practice trials using additional stimuli (e.g., *brushing*) preceded the main experiment. To make the color task comparable in difficulty to the gender task, images were desaturated using Photoshop software to a level of 3% (a level of saturation which made the color task more difficult but kept the actors distinguishable).

4.1.4. Data analysis

Data coding procedures and trial exclusion criteria were the same as in Experiments 1a and 1b. A mean of 14% (*SD* 4.9%) of trials (237 on average) per subject were excluded based on the previous exclusion criteria. Average accuracy was 96.2% (*SD* 2.7%), and average RT for the included data was 347 ms (*SD* 38 ms). Individual trial RTs were analyzed using linear mixed effects modeling with Event (repeated vs. switched), Side (repeated vs. switched), and Role (repeated vs. switched) as factors.

4.2. Results

As in Experiments 1a and 1b, a role switch cost was observed. In Table 3, we see that participants were on average 6 ms slower when the role of the target character changed from one trial to the next. This effect was again robust: 22/24 subjects and all 10 items went in the direction of the effect (Cohen's *d* of 1.42 and 1.40, respectively; see Figure 3). And although small, this effect was significant: The best-fitting mixed effects model included main effects of Role, Side, and Event, and interactions of Role × Side and Event × Side. The fit of the model was significantly better than the same model without the additional interaction of Role × Side, $\chi^2(1)$

[5] For Experiments 2 and 3, one extra repetition for each image stimulus (e.g., *kick-blue-left* → *kick-blue-left*) was included in case we found a need to examine exact image repetitions, but these were discarded a priori before analyses.

= 4.03, p =.04; and significantly better than a model that did not include Role at all, $\chi^2(2) = 31.9$, $p < .001$. Additionally, a model that also included an interaction of Role × Event was not a significantly better fit, $\chi^2(1) = 1.22$, $p = .27$.

Interestingly, the role switch cost interacted with repeated side, such that the role switch cost was greater when the side repeated than when it did not; importantly, however, the role switch cost remained even when the side was different. Like the additional effects observed in Experiments 1a and 1b, participants were faster when the side repeated, but only when the event category repeated. See Table 3 for a summary of the effects from the best-fitting model.

To summarize, a role switch cost was once again observed, even when the stimuli, event categories, and task were different. In fact, several participants reported that they explicitly tried to ignore the action as part of their strategy in performing the color task, but nevertheless, nearly all participants demonstrated the role switch cost. The results from this experiment suggest that the role switch cost is a general and robust phenomenon.

Table 3: Mean RTs across subjects for Experiment 2, separately for all factors that were significant in model fitting (significant interaction terms split by each factor level). 95% confidence intervals in parentheses.

Condition	Reaction time (ms)		Switch cost (ms)	*t* value for parameter in best-fitting model
	Repeated	Different		
Role	344 (16.2)	350 (16.4)	6 (1.75)	4.69*
Event	350 (16.6)	347 (16.2)	-3 (2.10)	-2.76*
Side	346 (16.1)	348 (16.9)	2 (6.00)	0.08
Role, Repeated Side	343 (16.0)	349 (16.3)	6 (2.41)	7.04*
Role, Switched Side	346 (16.9)	351 (17.0)	5 (1.89)	4.69*
Side, Repeated Event	344 (15.9)	353 (17.5)	9 (7.38)	2.60*
Side, Switched Event	346 (16.1)	348 (16.9)	2 (6.05)	0.08

* $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 4.2 for details on model comparisons.

4.2.1. Does Agent saliency drive the role switch cost?

Although the findings thus far provide evidence for a role switch cost, such a cost could be driven solely by a switch from Agent to Patient or vice-versa (i.e. it could be asymmetrical). Indeed, Agent primacy and saliency effects have been observed in both the linguistics and vision literature: Agents tend to precede Patients in linguistic utterances (Dryer, 2013; Goldin-Meadow, So, Ozyürek, & Mylander, 2008), and in continuous event perception, Agents attract attention, likely because they initiate movement before Patients (Abrams & Christ, 2003; Mayrhofer & Waldmann, 2014; Verfaillie & Daems, 1996) or because active body postures direct spatial attention (Freyd, 1983; Gervais, Reed, Beall, & Roberts, 2010; Shirai & Imura, 2016).

If Agent saliency is driving the role switch cost, we should observe two additional effects in our data across experiments: (1) different average RTs on trials in which the target was the Agent (Agent judgment trials) as compared to trials in which the target was the Patient (Patient

judgment trials); and (2) an asymmetry in the role switch cost, such that the cost for an Agent→Patient switch should be different from the cost for a Patient→Agent switch. Note that the directionality of the predictions (i.e. whether Agent trials should be faster or slower) depends on different theories about the interaction between event perception and the building of event structure. Under the view that Agents attract attention because of their active posture or movement initiation (e.g., Gervais et al., 2010; Verfaillie & Daems, 1996), one would predict faster RTs to Agent trials relative to Patient trials, since the primary task of participants was to locate the target actor. Under the view that observing Agents triggers the building of an event structure (Cohn & Paczynski, 2013; Cohn, Paczynski, & Kutas, 2017), attending to Agents (i.e. Agent judgment trials) might result in an additional cost due to initiation of event structure building, and therefore slower RTs. The crucial point here is that for Agent saliency (whether faster or slower) to explain the role switch cost, an asymmetry should also be observed between Agent→Patient and Patient→Agent switch trials, not only a difference between Agent and Patient judgment trials.

To formally test for these effects, we ran new mixed effects model comparisons in which we added Trial Judgment (Agent or Patient judgment trials) to the best-fitting models described in the above Results sections, separately for each experiment. Differences between Agent and Patient trials would manifest as a main effect of Trial Judgment, and an asymmetry in the role switch cost would manifest as an interaction of Role × Trial Judgment.

For Experiments 1a and 1b, adding a main effect of Trial Judgment or a Role × Trial Judgment interaction to the previously best-fitting models did not offer a significant improvement (all p's > .11). For Experiment 2, adding a main effect of Trial Judgment did significantly improve the fit over the previous best-fitting model ($\chi^2(1) = 55.5$, $p < .001$): Agent trial RTs were *slower* than Patient trial RTs (349 ms vs. 345 ms in subject means; see Table 4). The slower Agent RTs in Experiment 2 are in line with the hypothesis that Agents may trigger the process of "event building" (Cohn & Paczynski, 2013; Cohn et al., 2017). However, adding an additional interaction of Role × Trial Judgment to this model was not a significant improvement ($p > .66$). Given that differences between Agent and Patient trials was not consistent across experiments and that an asymmetry was not observed, these analyses suggest that Agent saliency cannot account for the role switch cost observed in the previous experiments.

Table 4: Mean RTs across subjects for each experiment, split by Trial Judgment type (Agent and Patient judgment trials, i.e. whether the target actor was the Agent or the Patient on each trial). 95% confidence intervals in parentheses.

Experiment	Reaction time (ms)		Agent trial advantage (ms)	*t* value for parameter in best-fitting model
	Agent trials	Patient trials		
Exp 1a (Gender Search, Catch Task)	383 (15.3)	383 (13.7)	0 (2.59)	0.58
Exp 1b (Gender Search)	387 (20.6)	387 (19.6)	0 (2.27)	1.58
Exp 2 (Color Search)	349 (16.3)	345 (16.2)	-4 (1.77)	-7.45*
Exp 3 (Color Search, Mirror-Flipped)	353 (24.7)	362 (23.1)	9 (2.91)	15.9*

* $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 4.2.1 for details on model comparisons.

4.3.
Discussion

Experiment 2 replicates and extends the findings from Experiments 1a and 1b by showing that role switch costs can be observed in explicit tasks other than those involving judgments about gender. Thus, these effects seem to be quite general.

5.
Experiment 3

In a final experiment, we probed the level of representation at which the role switch cost operates, testing two non-mutually exclusive possibilities. The first possibility, and the one of central theoretical interest to our investigation of event roles, is that the cost operates at the *relational level*: Agent and Patient roles are fundamentally relational (an Agent acts on a Patient), so perhaps it is the roles that scene entities take in an *interactive relationship* that results in the role switch cost. An alternative possibility, however, is that the role switch cost operates at the *pose level*: active body postures are probabilistically associated with Agents and not Patients (Hafri et al., 2013), so perhaps observed switch costs merely reflect salient changes in posture of the actors. Note that effects of posture, if they contribute to the switch cost, should have an equal effect whether the actors in the scene are interacting or not.

To test these two possibilities (*pose and relational levels*), we ran the same paradigm of Experiment 2, with one change: images were edited such that the actors always faced opposite directions ("mirror-flipped"). With this manipulation, the actors' poses were preserved but their interaction was substantially reduced or eliminated (see also Glanemann et al., 2016). Thus, any switch costs observed in the current experiment (with non-interactive actors) can only be attributed to switches at the *pose level*.

The image manipulation in the current experiment will allow us to assess the specific contribution that two levels (*pose and relational levels*) make to the switch costs observed in our previous experiments. If the previously observed role switch costs were due only to informational conflict at the *relational level*, we should observe a complete elimination of the switch cost here, since any interaction between actors is now minimally present. If the switch costs were due only to the *pose level*, then there should be no consequence of the image manipulation: all and only the previous role effects should obtain. However, if the role switch cost in previous experiments was due to conflict at both levels (*relational* and *pose*), the switch cost should still obtain here, but its magnitude should be significantly lower than that of the switch cost in this experiment's closest counterpart (Experiment 2).

5.1.
Methods

5.1.1.
Participants

An additional 24 members from the University of Pennsylvania community participated and

received class credit. Given the stability of the role switch effect across Experiments 1a, 1b, and 2, we believed this number to be sufficient. Data from an additional four participants were excluded: two for not completing the experiment and two for low accuracy (<86%). This accuracy threshold was based on performance of participants in the previous experiments (all >89%), although inclusion of these excluded participants did not qualitatively change the results.

5.1.2. Materials and procedure

Stimuli from Experiment 2 were edited in Photoshop such that actors always faced away from one another. This was achieved by flipping each actor (or both) horizontally about his own center axis. Since actors sometimes partially occluded one another (e.g., in *slapping*, the Agent's hand and Patient's face), this procedure occasionally resulted in missing body or face parts in the images. The missing regions were replaced with parts from other images using various Photoshop tools. This was successful: no subject noticed the image manipulation even when questioned during debriefing. Example images appear in Figure 1, and examples for each event category appear in the Appendix. Apparatus and procedure were identical to Experiment 2.

5.1.3. Data analysis

Data coding procedures and trial exclusion criteria were the same as in Experiments 1a, 1b, and 2. A mean of 12% (SD 3.2%) of trials (190 on average) per subject were excluded based on the previous exclusion criteria. Average accuracy was 97.7% (SD 1.6%), and average RT for the included data was 358 ms (SD 56 ms). Main analysis procedures were the same as in Experiment 2. Although in principle the actors were no longer Agents and Patients due to the mirror-flip manipulation, we coded Role (repeated vs. switched) based on each actor's corresponding role in the unedited stimuli.

5.2. Results

A role switch cost was once again observed. In Table 5, we see that participants were on average 3 ms slower when the role of the target character changed from one trial to the next. This effect was robust here as well: 20/24 subjects and 8/10 items went in the direction of the effect (Cohen's *d* of 0.86 and 0.97, respectively; see Figure 3). And although small, it was significant: The best-fitting mixed effects model included main effects of Role and Side. The fit of the model was significantly better than the same model that did not include Role at all, $\chi^2(1) = 13.8$, $p < .001$. Additionally, a model that also included an interaction of Role × Side was not a significantly better fit, $\chi^2(1) = 0.10$, $p = .75$, nor was a model that also included a main effect of Event, $\chi^2(1) = 0.01$, $p = .92$. As in the previous experiments, participants were slower when side repeated. See Table 5 for details.

Table 5: Mean RTs across subjects for Experiment 3, separately for all factors that were significant in model fitting. 95% confidence intervals in parentheses.

Condition	Reaction time (ms)		Switch cost (ms)	*t* value for parameter in best-fitting model
	Repeated	Different		
Role	356 (23.5)	359 (24.2)	3 (1.59)	3.86*
Side	363 (26.1)	353 (22.0)	-10 (6.81)	-15.8*

* $p < .05$ in best-fitting mixed effects model (calculated using R lmerTest package). See section 5.2 for details on model comparisons.

5.2.1. Comparison of Experiments 2 and 3

Given that Experiments 2 and 3 are a minimal pair, they present an ideal opportunity for additional assessment of the contributions of the *pose* and *relational levels* to the role switch cost. Because of the mirror-flip manipulation in the current experiment, the role switch cost here can only be attributed to the *pose level* (since the interaction between actors was minimal or non-existent), while in Experiment 2 it can be attributed to both *pose* and *relational* levels. Indeed, the size of the standardized effect in Experiment 3 was about two-thirds of that observed in Experiment 2 (see Tables 3 and 5). To formally compare the role switch cost across experiments, we ran new mixed effects models with the data from both experiments, with a base model whose random effects structure, main effects, and interactions were identical to the best-fitting model of Experiment 3. Adding a main effect of Experiment and interaction of Role × Experiment to the base model significantly improved the fit as compared to a model with only a main effect of Experiment, $\chi^2(1) = 10.6$, $p = .001$. This comparison yields credence to the idea that a combination of levels (*pose* and *relational*) led to the switch costs observed in Experiment 2.[6]

5.2.2. Does Agent saliency mediate the role switch cost in this experiment?

Here, unlike in previous experiments, there was a reliable Agent trial advantage: participants were on average 9 ms faster to respond on Agent judgment than Patient judgment trials. This was confirmed in mixed effects models: adding Trial Judgment (Agent vs. Patient judgment trial) as a factor to the best-fitting model from above significantly improved the fit, $\chi^2(1) = 252$, p

[6] In the mirror-flip manipulation, it could be argued that the interactive nature of the actors is not completely eliminated; for example, a kicker facing away from a would-be kickee may appear instead to be marching away from the other actor – a kind of social interaction. If this is the case, the reduced effect here could be due to a reduction (but not full elimination) of the interaction between actors, rather than a combination of the relational and pose level information. However, based on responses to questions during debriefing, the majority of participants considered the actors non-interacting. Thus, although the role switch cost in this experiment should perhaps be called a "posture switch cost", we use the term "role switch cost" for consistency with the previous experiments.

< .001. Furthermore, this Agent advantage was greater than in any other experiment (independent samples *t* tests over subjects: all *t*'s > 6.40, *p*'s < .001; paired samples [Experiment 2] and independent samples [Experiments 1a and 1b] *t* tests over items: all *t*'s > 3.31, *p*'s < .01; see Table 4 for the magnitude of Agent advantage in each experiment). As discussed in section 4.2.1, for Agent saliency to account for the results here, we would also expect an asymmetry in the role switch cost, i.e. a differential cost for Patient-switch than Agent-switch trials. However, this additional effect was not observed: adding an interaction of Role × Trial Judgment did not improve model fit over a model with only a main effect of Trial Judgment, $\chi^2(1) = 2.02$, $p = .16$. Thus, we can conclude that Agent saliency (or more properly here, "active posture" saliency) did not mediate the role switch cost in the current experiment.

The contrast in directionality of the Agent saliency effects between Experiments 2 and 3 is further evidence that these stimuli were analyzed at different levels (*pose* vs. *relational*) by the participants in each experiment. In Experiment 2, Agent trials were slower than Patient trials, consistent with the hypothesis of Agents triggering event-building in visually analyzed event scenes due to Cohn et al. (2013; 2017). In the current experiment (Experiment 3), we speculate that a different process may be at work: the actors were analyzed at the postural level, with no event building initiated (given that actors in the scene were not interacting with one another). The robust effect of Trial Judgment (faster Agent judgment, or "active posture" trials) in this experiment is consistent with previous work that argues that active postures independently guide attention in scenes (Freyd, 1983; Gervais et al., 2010), even for infants (Shirai & Imura, 2016).

5.3. Discussion

We again observed a reliable role switch cost, but this differed substantially from our previous experiments. First, the effect size here was roughly two-thirds that of Experiment 2. Second, unlike in previous experiments, an Agent (active posture) advantage also obtained. Thus, the *pose level* alone (i.e., active and passive posture differences associated with certain roles) cannot account for the entirety of the role effects across studies. Instead, the role switch cost observed in previous experiments was likely operating at both the *pose* and *relational levels*.

Given the differences observed between Experiments 2 and 3, we propose that the perceptual system may be differentially attuned to interacting and non-interacting individuals. On the one hand, the perceptual system is likely tuned to active postures generally, in line with evidence that active body postures direct spatial attention (Freyd, 1983; Gervais et al., 2010; Shirai & Imura, 2016). But for interactive events (Experiments 1a, 1b, and 2), we hypothesize that attention naturally spreads to both actors (the Agent and Patient). Indeed, recent work has shown a facilitatory effect on recognition of two-person interactions (relative to non-interacting dyads) akin to the well-known face-inversion effect, such that inversion effects are found for stimuli in which two people are facing each other but not when they are facing away (Papeo, Stein, & Soto-Faraco, 2017). Although our experiment was not explicitly designed to test for a general attentional advantage for interacting vs. non-interacting actors, we did find some evidence that such an advantage may exist. RTs were approximately 11 ms lower in Experiment 2 (in which actors were in interactive relationships, mean RT 347 ms) than in Experiment 3 (in which actors were mirror-flipped, i.e. not interacting, mean RT 358 ms). This was confirmed in

a paired *t* test comparing RTs for individual image stimuli across the two experiments, collapsing over all cross-trial switch factors (e.g., the mean inverse RT for the image of *blue-kicking-red-from-the-left* in Experiment 2 compared to its mirror-flipped equivalent in Experiment 3), $t(39) = 11.5$, $p < .001$, $d = 1.83$.

Given that accuracy on the main task (color search) was actually numerically *higher* in Experiment 3 vs. Experiment 2 (97.7% vs. 96.2%, respectively), we do not believe the overall RT difference between the two experiments is due to general confusion on account of the mirror-flip manipulation; instead, the RT difference supports the hypothesis that there is an attentional advantage specific to interacting human figures, as if the perceptual system treats the interacting figures as an attentional unit.

5.3.1. Can the role switch cost be attributed to order effects or to the large number of trials used?

One general concern across experiments is that – although the large number of trials per subject (about 1600) resulted in robust estimates of central tendency – we might be capturing an effect that is due to the peculiarities of the task. This could surface as order effects: perhaps the role switch cost is due to effects of getting acquainted with the task (gender or color search), or perhaps it is an effect that emerges from overlearning the response to each stimulus, or to fatigue. We tested these possibilities directly, by adding additional interaction terms for Role (the switch cost) and either Trial Number (1 to approx. 1600) or Block Number (1 to 40) to the best-fitting model for each experiment. Adding the Role × Trial Number interaction term did not improve any of the model fits, all $\chi^2(1) < 1.64$, p's > 0.20, nor did adding the Role × Block Number interaction term (with an additional main effect of Block Number), all $\chi^2(1) < 1.47$, $p > 0.23$. Thus, it seems unlikely that the role switch cost is driven by any peculiarities attributable to order effects, such as gradual accommodation to the task, overlearning, or fatigue.

Additionally, given that we obtained such a large number of observations per subject (about 1600), we wanted to ask whether we would have observed the role switch cost with fewer observations than were obtained in each experiment. To test this, we performed a power analysis that tested at which point in the experiment, if we had stopped collecting data, we would have found a significant role switch cost (at a standard significance level of $\alpha = .05$). Specifically, separately for each experiment, we performed identical mixed model comparisons to those reported in each experiment above, using the same best-fitting models (i.e., comparing the likelihood ratio values for models with and without Role as a factor). This was performed on data from each block, cumulatively (e.g., for Cumulative Block 1, this only included data from block 1; for Cumulative Block 2, data from both block 1 and 2; for Cumulative Block 3, data from blocks 1-3; etc., all the way up to block 40, which included data from the entire experiment). We simply asked at which block significance ($p < .05$) was reached and maintained for subsequent blocks in model comparisons. This is depicted in Figure 4. We find that for Experiments 1a and 2, as little as one-tenth of the data was sufficient to reach and maintain significance, and for Experiments 1b and 3, about half to two-thirds. Thus, we can be confident that in general, our estimate of the amount of data required was conservative, and we likely would have detected the

role switch cost even with many fewer observations per subject and item.[7]

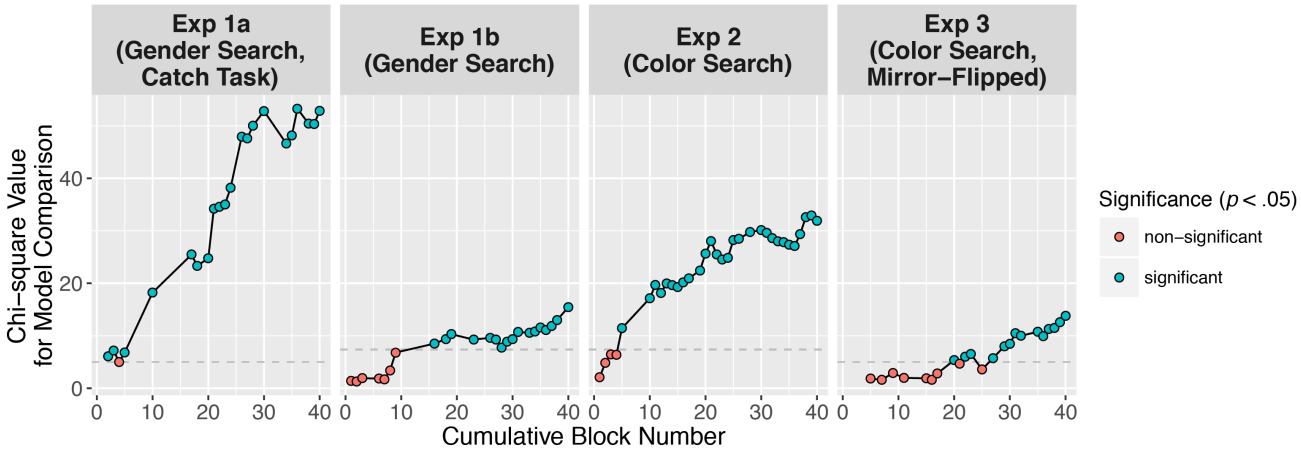


Fig. 4. Analysis of the amount of data required to obtain a significant role switch cost effect in each experiment. Mixed effects model comparisons (for models with and without Role as a factor) that were identical to those reported for each experiment were calculated on data from each block, cumulatively (i.e., for *cumulative block number* on the x-axis, each block number also includes data from all previous blocks, e.g., the data point for block number 30 represents a statistic calculated using models with data from blocks 1-30). The dotted line in each plot indicates the chi-square value required for a level of significance of $p < .05$ for that experiment's model comparison. Blue points indicate significant chi-square values. (Points for some data subsets do not appear because models using those subsets did not converge). These plots indicate that fewer blocks of trials would have been sufficient for detecting the role switch cost in each experiment (in some cases, such as in Experiments 1a and 2, we would have detected the switch cost with as little as one-tenth of the data).

5.3.2. Can linguistic encoding of our stimuli explain the role cost?

Given that there is evidence of rapid interplay between event apprehension and utterance formulation (Gleitman, January, Nappa, & Trueswell, 2007), it is conceivably possible that linguistic encoding of the stimuli was happening, even within this short time frame (<400 ms). If the switch cost we observed is due to purely *grammatical* categories (Subject, Object), then our experiments cannot adjudicate the generality of event roles (i.e., Agent and Patient, or related cluster-concepts; Dowty, 1991; White et al., 2017). In other words, *kicker* and *tickler* may not be conceptually related, but when they are situated in utterances, the *kicker* and *tickler* become similar by virtue of their both being grammatical Subjects (the same reasoning applies

[7] We should note that the experiments reported in this manuscript were the first that we conducted using this switch cost paradigm, and the first (to our knowledge) to use this method in scene perception research in general. Therefore, given our initial uncertainty in how strong of an effect we should observe in such a paradigm, we used a large number of trials per subject to maximize our chances of observing an effect of event role if it were to exist. Since we found that only a subset of the trials was needed to detect the role switch cost in our experiments, we hope that the reported power analysis proves useful to other researchers interested in using a similar paradigm for asking questions about encoding of event information in visual scenes.

to *kickee* and *ticklee*).

However, linguistic encoding is unlikely to explain the role switch costs observed in our experiments for several reasons. First, explicit linguistic encoding was rare: in post-experiment questioning, only nine subjects across all experiments reported linguistically encoding the stimuli at any point in terms of who did what to whom (2 in Experiment 1a, 5 in Experiment 1b, 2 in Experiment 2, and 0 in Experiment 3). Second, any linguistic encoding that occurred appears to have had little influence on the role switch cost: the cost was not statistically different between participants that reported encoding the events linguistically and those that did not, for any experiment (all p's > 0.20, unpaired t-tests). In fact, only two of the nine participants that reported linguistic encoding, both in Experiment 1b, appeared in the top 50th percentile of switch cost magnitude among the other participants in their experiment.

It is also unlikely that participants were linguistically encoding the events implicitly. If they were, then we might expect a grammatical Subject advantage: Subjects appear first in utterances in English (a Subject-Verb-Object language), so trials on which the target actor was the Agent (the grammatical Subject in canonical active-voice utterances) might show faster RTs than when the target actor was the Patient (the grammatical Object). However, this was not the case: Agent (Subject) trials were actually significantly *slower* than Patient (Object) trials in Experiment 2, and there was no significant Agent (Subject) advantage in Experiments 1a and 1b (see Table 4).

Taken together, these analyses suggest that – although some participants did report encoding the stimuli linguistically – it had little if any influence on the role switch effects observed in our studies. Future work could further probe the influence of language on performance in a task such as ours by testing participants from different language groups, or those without access to fully formed natural language (e.g., deaf homesigners; Feldman, Goldin-Meadow, & Gleitman, 1978; Zheng & Goldin-Meadow, 2002).

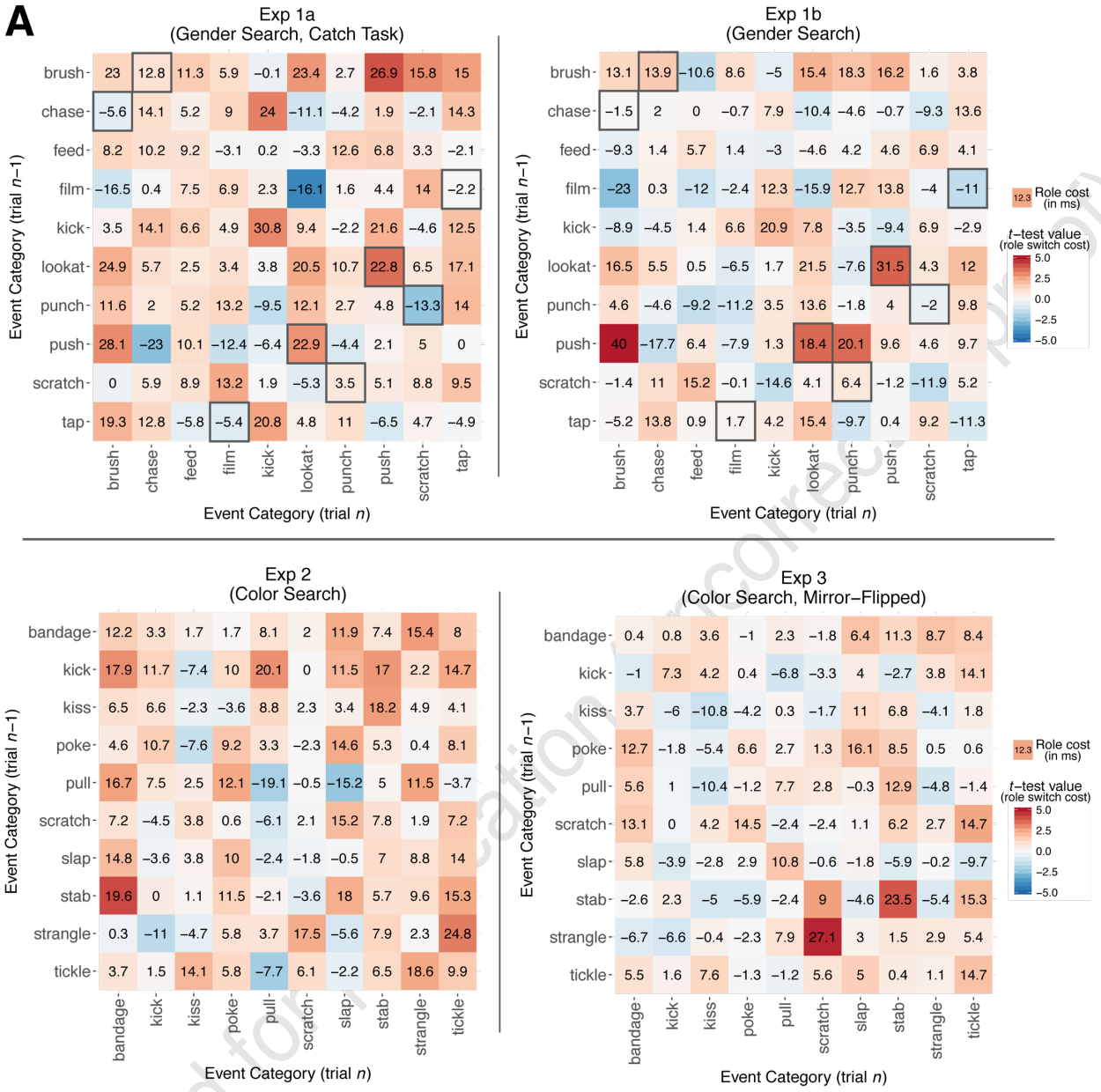
5.3.3.
How general is the role switch cost over transitions between particular event categories?

In previous analyses, we found some evidence that the role switch cost is at least partly event-general (i.e. not tied to the specific preceding event category): in Experiments 1a and 1b, the role switch cost still held when Actor Pair (and therefore Event Category in that stimulus set) switched (see Tables 1 and 2 and section 3.2.1); and in Experiment 3, there was not a significant interaction of the role switch cost with repeated/switched event category. However, it still could be the case that the cost is dependent on which particular event categories precede others (i.e. that the role switch cost is driven by a small subset of preceding event categories). For example, in the extreme, it could be that the role switch cost for each event category is obtained only when preceded by the category *kicking*.

To address this, we simply calculated the average role switch cost (using inverse RTs) across subjects for each event category to every other event category, collapsing over Agent side. This yielded a 10 × 10 matrix of values for each experiment, where each cell of a matrix represents the average role switch cost for a transition from one particular event category to another, illustrated in Figure 5A (using raw RTs). We then tested whether these event-to-event role switch costs were significantly above zero for each experiment. Indeed as illustrated in Figure 5B, this was the case (all $t(99) > 2.39$, $p < 0.02$), even when excluding transitions between the

same event categories, i.e. the diagonals of the matrices (all $t(89) > 2.03$, $p < 0.05$).[8] These analyses suggest that, at least for the event category exemplars used in our experiments, there is some commonality across the roles of the participants in different event categories that is driving the role cost. Implications for event role representations more broadly appear in section 6.1.

[8] The same analyses can be conducted using mixed effects models, testing whether the effect of Repeated Role no longer significantly improves model fit once event-to-event transitions are taken into account (operationalized here as separate random intercepts for Previous Event and the Previous Event × Current Event interaction, with random slopes for Repeated Role for each random intercept). These analyses support the same conclusion as the *t*-test analyses in the main text, namely that the role switch cost is not driven by a small set of event category transitions.



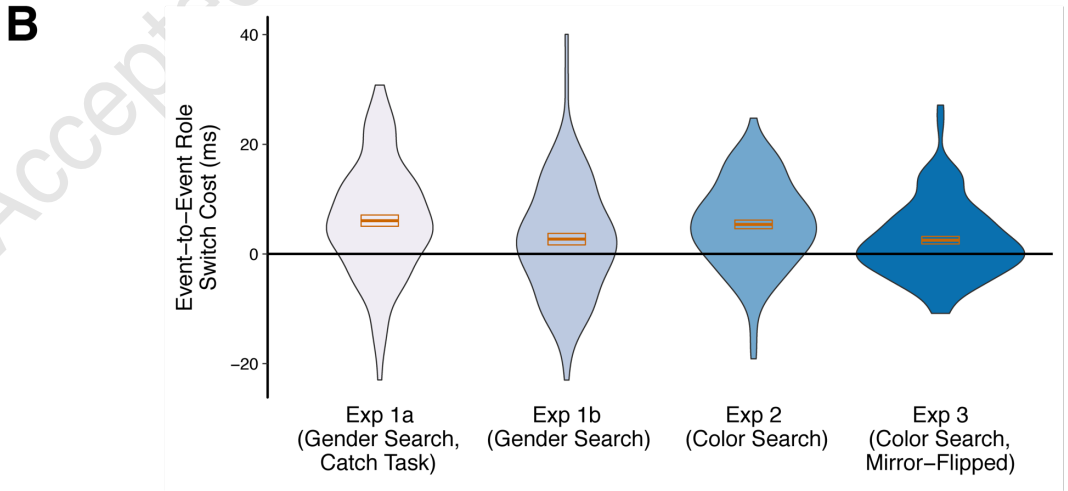


Fig. 5. (a) Mean role switch cost over subjects (in milliseconds) calculated between each Event Category and every other Event Category, collapsing across Actor Side, separately for each experiment. Color shading indicates *t*-test values for the switch cost across subjects ($|t(23)| > 2.07$ is significant at $p < .05$ uncorrected), with red indicating a role switch cost, and blue a role switch benefit. Gray boxes around cells in Experiment 1a and 1b matrices indicate transitions between different Event Categories that feature the same Actors (see section 2.1.2), which was found in analyses to result in higher switch costs (see section 3.2.1); this is not indicated in Experiments 2 and 3 since there was always only one set of actors. Note that diagonals in each matrix represent the switch cost for the same Event Category, so always reflected the same set of actors, in all experiments. (b) Violin plots of all cells from the four matrices in (a). Violin plot outlines indicate the kernel probability density, i.e. the width of each plot indicates the proportion of event-to-event transition values at each role cost magnitude. Orange boxes indicate the mean and standard error across transition values, for each experiment. Analyses showed that the role switch cost was not driven by a small subset of event-to-event transitions: as can be seen, the majority of values were above zero.

6.
General Discussion

Our experiments demonstrate that the structure of an event, i.e. who acted on whom, is spontaneously encoded in visual processing, even when attention is directed toward other visual features (here, gender or color). This process manifested as a role switch cost, i.e., a relative lag in reaction time when the role of the target actor switched from trial to trial. The effect was robust across stimuli, event categories, and task (Experiments 1a and 1b: gender search; Experiment 2: color search). In Experiment 3, we determined that the role switch cost observed in the previous experiments cannot be fully explained by body posture differences associated with Agents and Patients. Furthermore, we found that the cost was not driven by a subset of the possible transitions from one event category to another, suggesting that the role information computed is quite general. Taken together, our experiments demonstrate (for the first time, to our knowledge) both the rapidity and generality of the event role computation itself.

6.1.
Implications for event role representations

Although we have shown that assignment of Agent and Patient to entities in visual scenes is rapid and spontaneous, it may be that in continued processing, this coarse role assignment can be reversed or refined, in at least three ways. The first is additional visual input, in the form of successive fixations: for example, upon further observation, perhaps one recognizes that the initially identified Patient is holding a weapon, making him an Agent (a *shooter*); or that an Agent is holding an object to transfer, making the Patient a Recipient. Indeed, a recent gist-extraction study of event scenes revealed that observers need substantially longer viewing times to identify the coherence of spatially local event properties such as the category of instrument objects vs. global event properties such as posture/orientation (Glanemann et al., 2016). The study of Glanemann et al.(2016) highlights the advantage afforded by initial commitment to a coarse role assignment: it can help guide scene fixations in a targeted manner (see also Castelhano & Henderson, 2007).

A second way that role assignment can be reversed or refined is via flexible event construal:

Despite how an event plays out in the world, people can construe it in an innumerable number of ways (sometimes for comedic effect: "Yeah, I'm fine. I snapped my chin down onto some guy's fist and hit another one in the knee with my nose"; Ross, 1972). We speculate that in general, flexibility in event construal reflects a top-down, cognitive re-interpretation of an initial commitment provided rapidly by the visual system.

Finally, the context in which an event occurs likely allows for later assignment of more event-specific roles like *helper* or *hinderer* that incorporate this contextual information. Indeed, there is developmental evidence for both event-general and event-specific role distinctions: young infants readily distinguish Agents and Patients in social events like *helping* and *hindering*, but they also themselves prefer positively valenced Agents (i.e., *helper*; Hamlin, Wynn, & Bloom, 2007; Kuhlmeier, Wynn, & Bloom, 2003).

Given that in our experiments, we found the role switch cost to be somewhat event-general, an important theoretical question is whether there are *systematic* differences in the role switch cost in terms of hypothesized properties of roles in different event categories. In particular, some theories of event roles hypothesize that certain components of events (e.g., contact, causation, and change of state or motion) are conceptual primitives, posited as such because they are relevant for grammar (i.e., they constrain the sentence frame in which a verb can be used; Levin, 1993; Levin & Rappaport-Hovav, 2005; Pinker, 1989; Talmy, 2000) or because they are available early on in development (Strickland, 2016). Notably, these event components are similar to features proposed in cluster-concept notions of event roles (Dowty, 1991; Kako, 2006; White et al., 2017).

Although the consistency we observed in the role cost across events is broadly suggestive of generality (see Figure 5, and section 5.3.3), we do not believe we have a convincing way to address the precise characteristics of this generality with the current data, for the following reasons. First, the event categories we used did not independently vary in theoretically relevant event components such as cause, contact, state-change, and motion. Second, we had essentially only one exemplar (i.e. one postural "tableau") per event category (see Figure 1 for examples). Thus, to address the generality and granularity of event roles extracted from visual scenes, future work will need to include many more event categories and to systematically manipulate hypothesized event components within event category.

Whatever theoretical distinctions end up accounting for the complexities of an observer's event conceptualization, we assert that there is a rapid and spontaneous assignment of Agent-like and Patient-like roles to interactive event participants, possibly before more refined role distinctions (e.g., Recipient) or social contingencies (as in the *helping/hindering* case) have been computed, and in some cases before event-specific role identification occurs (e.g., *kicker, kickee*).

Consequently, now that we have established the robustness and generality of the basic phenomenon of spontaneous role extraction with Agent-like and Patient-like event participants, there is a large set of theoretically interesting questions about how the visual system parses the roles in events with different numbers of participants and different relationships among them. For example, in single-participant events where the participant undergoes a change of state or location (e.g., *melting*, *falling*), is the participant assigned a Patient-like rather than Agent-like status? In a joint interaction such as *dancing*, are participants may be assigned similar roles (e.g. both Agents) rather than Agent and Patient? What is the role status of participants in complex events such as transfer events (e.g., *giving*, *serving*)?

6.2. Implications for the relationship between perceptual and linguistic encoding of event roles

The early stages of event perception as examined in the current studies have the potential to inform theories of argument selection in linguistic descriptions of events (i.e., whether event participants belong in sentential subject, object, or oblique positions). Our general theoretical viewpoint consists of the following notions: (1) in early perceptual processing, scene entities are categorized as Agent-like and Patient-like, often before the event category itself is determined; and as such, (2) initial role categorization is not dictated primarily by the event category itself (along with the corresponding verb-specific roles such as Stimulus Experiencer, and Instrument), but rather by the perceptual particulars of the scene, i.e. the particular *token* of the event category. Our studies provide support for these notions: we found role switch costs even across exemplars of event categories that would not be considered in the literature to be canonical Agent-Patient relationships: events with a mediating instrument (*stab*, *film*, and *bandage*); events without caused motion or state-change (*look at*, *call after*, and *film*); and an event of transfer (*feed*), where the Patient might more traditionally be considered a Recipient.[9] Our viewpoint provides a possible perceptual explanation for at least two issues in linguistic argument selection: (1) the optionality and argument status of some event participants, such as Instruments; and (2) the cross-linguistic variability in grammatical status of certain event roles, such as Stimulus and Experiencer.

First, let us consider the optionality and argument status of event participants. It is debated whether instruments should be considered arguments of verbs: to describe a *stabbing* event, for example, one may say *John stabbed the man* or *John stabbed the man with a knife*. Rissman and colleagues (2015) account for these inconsistencies at the level of event construal: argumenthood depends on construal of a *particular token* of an event as indicated by a verb and its sentential context, rather than an absolutist notion of arguments that depends solely on the verb itself. Our work provides a perceptual complement to this notion: we argue that early available perceptual cues to role assignment have a strong influence on initial event construal. Hence, the degree of perceptual salience of objects involved in a particular token of an event should partially determine the degree to which an argument of a verbally encoded event will be optional, or should be considered an argument at all in the case of Instruments (see also Brown & Dell, 1987, on the pragmatics of inclusion of event participants in discourse).

The rapid and spontaneous encoding of event participants as Agent-like and Patient-like might also account for the fact that linguistic argument selection for certain event categories is more consistent cross-linguistically than for others. For example, the Agent- and Patient-like

[9] Of course, the scene exemplars (the images used for *look at*, *feed*, etc.) were selected precisely because there *was* general agreement in our previous study (Hafri et al., 2013) about the roles of the scene participants (who was performing the action vs. being acted upon). However, the fact that we found the role cost even for these items suggests that it is in principle possible to find Agent and Patient-like role effects even for categories of events without canonical Agent-Patient relationships. This provides evidence that the category of event does not exert a strong influence on early role assignment.

status of the subject and object in a description of a *hitting* event is fairly straightforward. In contrast, the statuses of subject and object in a description of a *frightening* or *fearing* event are much less clear (e.g., *John frightens Mary* and *Mary fears John* can describe the same event; Dowty, 1991), with some hypothesizing thematic roles distinct from Agent and Patient for these event participants (i.e., Stimulus or Experiencer, dependent on which participant is seen as the implicit cause of the event; Hartshorne, 2014; Levin & Rappaport-Hovav, 2005). We hypothesize that from instance to instance of a given event category, the Agent- and Patient-like perceptual properties of the participants may on average be less variable (e.g., *hitting, kicking*) or more variable (e.g., *fearing, frightening, looking*). Thus, it is not surprising that event categories involving Stimulus/Experiencer-like roles (e.g., *fearing*) are the ones for which there is high cross-linguistic variability in terms of which participant must appear in subject position. Indeed, we have previously argued that the high degree of cross-linguistic correspondence between Agents/Patients and subjects/objects is probably not a coincidence, but rather reflects a fundamental relationship between "core" cognition and perception (Strickland, 2016).

This brings us to the question of the degree to which language dictates conceptual event role assignment. It has certainly been shown that the linguistic framing of an event may influence attention to and memory for certain event participants or event components (e.g., Fausey, Long, Inamori, & Boroditsky, 2010; Kline, Muentener, & Schulz, 2013; Papafragou, Hulbert, & Trueswell, 2008; Trueswell & Papafragou, 2010). Notice here, however, that these phenomena reflect how language production alters attention in scenes ("looking for speaking"), or how language comprehension affects event construal (serving as a marker of the scene entities considered relevant by the speaker). We predict that cross-linguistic differences should be minimal in the first moments of event perception, and only afterward might language-specific effects be observed, if at all (e.g., language-specific conventions in terms of assignment of Stimulus and Experiencer to certain grammatical positions). Such a prediction could be tested by running our experimental paradigm with speakers of different languages, or with populations with no exposure to fully formed natural language, e.g. deaf homesigners (Feldman et al., 1978; Zheng & Goldin-Meadow, 2002). A second prediction is that an observer's event construal will be more susceptible to linguistic modulation when the ambiguity of the initial role information available in the scene is higher, such as with Stimulus-Experiencer events, e.g. *frightening*, where the relative Agent-like or Patient-like cues between event participants may not significantly differ. In other words, speakers certainly use specific verbs and frames in an event description to convey the importance of the various event participants to their event construal (e.g., *frighten* vs. *fear*), but an observer's construal depends heavily on the perceptual parameters of the interaction *in the first place*.

To summarize this section, we believe that our results help to address some puzzles in the linguistic encoding of events, such as the argument status of event roles like Instruments and the cross-linguistic variability in grammatical status of certain roles like Stimulus and Experiencer. We speculate that in the first moments of event perception, how Agent-like and Patient-like scene participants are, as well as their perceptual salience, matters more for event construal and subsequent linguistic encoding than the logical relationship between event participants (such as Stimulus/Experiencer) in the depicted event category.

6.3. Implications for high-level visual perception

Our work is consistent with a wealth of previous literature that has demonstrated rapid, bottom-up encoding of semantic content from visual scenes (Biederman et al., 1982; Castelhano & Henderson, 2007; Greene & Fei-Fei, 2014; Greene & Oliva, 2009; Potter, 1976; VanRullen & Thorpe, 2001). Crucially, we find that not only are the perceptual features that are correlated with event role (i.e., body posture) extracted by the visual system rapidly, but the computation of the abstract role information itself is rapid. Observers in our studies viewed the scenes for less than 400 ms (based on mean response times), so for us to have obtained the role switch cost, the computation of role information must have taken place within this time frame.

Our findings fit within a broader literature in visual perception which shows that spontaneous and possibly automatic perceptual processes are not limited to low-level properties (e.g., lines and edges), but also extend to "high-level" representations that include objects (Scholl, 2001), event types (Strickland & Scholl, 2015), causality (Kominsky et al., 2017; Rolfs, Dambacher, & Cavanagh, 2013), and animacy (van Buren, Uddenberg, & Scholl, 2015). Like our event role results, these other processes often map neatly onto representations from the literature on infant "core cognition" and potentially conflict or diverge from higher-level, explicit judgments (Cheries, Wynn, & Scholl, 2006; Spelke & Kinzler, 2007; see Strickland, 2016, for a discussion of the relationship between elements of "core" cognition and cross-linguistic grammatical patterns). Additionally, the differences in the role switch cost for interactive actors (Experiment 2) and non-interactive actors (Experiment 3) supports the hypothesis that another element of core cognition that is reflected in perception are the social interactions of others, including their roles (Spelke & Kinzler, 2007). This is in line with other recent work suggesting that the perceptual system treats interacting figures as an attentional unit (Papeo et al., 2017) and that there is a region in the human brain selective for observed social interactions (Isik, Koldewyn, Beeler, & Kanwisher, 2017).

An open question is the extent to which the role switch cost is specific to human interactions, or is a reflection of more general processing of the interactive relationships between scene entities, both animate and inanimate. That is, in event scenes that involve interactions with or among inanimate objects (e.g., a woman opening a door or a ball hitting a rock), are roles assigned using similar visual processes? Given our assertion that early in visual processing, scene entities are assigned coarse Agent-like and Patient-like roles, it follows that, if an inanimate object is salient enough in the visual representation, it should also be rapidly assigned an Agent-like or Patient-like role. However, there is evidence that visual processing of animate and inanimate entities is quite distinct, both in terms of differential attention (Kominsky et al., 2017; van Buren et al., 2015) and underlying cortical pathways (Connolly et al., 2012; Scholl & Gao, 2013). It will require further investigation to determine whether the visual system assigns roles similarly to animate and inanimate scene entities.

6.4. Implications for action and event perception

Researchers studying action perception and its neural substrates have tended to focus on single-actor actions (e.g., *walking*; Giese & Poggio, 2003; Lange & Lappe, 2006) or actor-object

interactions (e.g., *grasping*, *opening*; Rizzolatti & Sinigaglia, 2010; Wurm & Lingnau, 2015 and many others). Our work suggests that to gain a complete picture of action perception and the neural substrates supporting it, researchers must also study the event structure of actions and interactions (Hafri et al., 2017). Additionally, our results have implications for theories of event perception from ongoing activity, particularly Event Segmentation Theory (EST; Zacks, Speer, Swallow, Braver, & Reynolds, 2007). EST holds that during continuous perception, people construct an "event model" that includes relevant causes, characters, goals, and objects (Zacks, Speer, & Reynolds, 2009). Importantly, EST implies that this process does not require conscious attention. Our results directly support this core implication: people rapidly and spontaneously encode the structure of observed events, even when attention is guided to other properties of observed scenes. Our results further suggest that event roles should be considered key components of event models themselves, an intuitive notion: if event roles change, then so does the currently observed event.

6.5. Spontaneity vs. automaticity of role encoding

In the introduction to this paper, we defined a spontaneous process as any process that is executed independently of an explicit goal. Such a process could be automatic, in the sense that it is mandatory given certain input characteristics, but it could also be spontaneous but not automatic in the sense that, under some conditions and with some cognitive effort, the process could be prevented from being executed. Our results at minimum demonstrate the spontaneity of role encoding. However, what can we say about the potential automaticity of event role encoding?

One criterion for automaticity is the notion of "ballistic" engagement, i.e. that given certain types of perceptual input, a particular process is necessarily engaged and runs to completion (e.g., an English speaker cannot help but process the sounds of English as such; Fodor, 1983). Additional criteria are due to Shiffrin and Schneider (1977), who studied target item search among distractor items: they assert that automatic processing is quick, is not hindered by capacity limitations of short-term memory, and requires only limited attention. One difficulty in assessing the degree of automaticity using these criteria is that there is not a straightforward mapping between Shiffrin and Schneider's definitions of target and distractor and our definitions in the present study. In Shiffrin and Schneider (1977), targets and distractors are different objects (e.g., letters and numbers) on screen. In contrast, in our paradigm, the "target" (gender/color information) and "distractor" (role information) are two levels of description of the *same entity* (the target actor). Thus, if attention to *different levels* of the same stimulus and to *different stimuli* should be considered analogous under the Shiffrin and Schneider criteria, then our results are consistent with automaticity: even when attention is directed to one level of the target actor (gender/color), we find that subjects also encode the same entity at another level (role).

However, since gender and color in our stimulus set were not in direct conflict with role information, only orthogonal to it, answering whether role extraction is automatic rather than simply spontaneous requires further research. Notably, such a distinction between spontaneity and automaticity is relevant not only within the domain of the current study, but applies to

many fields investigating processes that have the potential to be considered automatic (e.g., theory of mind; Leslie, 1994; Scholl & Leslie, 1999).

6.6. Practical vs. theoretical significance of the role switch cost

Before we close, we believe that a separation of the empirical robustness, practical consequences, and theoretical import of the role switch cost is warranted. The empirical evidence is clear. We have reported a highly replicable effect, with each experiment showing a consistently large standardized effect size (minimum Cohen's *d* 0.55), and with a majority of subjects and items showing the effect in all cases. We also demonstrated that the large number of observations per subject were not necessary to obtain the effect (see Figure 4 and section 5.3.1).

For practical purposes, we are not surprised at the small absolute magnitude of the effect (about 5 milliseconds), since our experiments were explicitly designed to dis-incentivize people from making role categorizations. Remarkably, even under these fairly extreme conditions, participants exhibited a trace of tracking event roles. Nevertheless, we would expect whatever mental mechanisms that produce the tiny absolute effect sizes here to matter much more in everyday situations where Agency and Patiency *are* task-relevant (e.g. for the purposes of producing language or judging the behavior of conspecifics).

We assert that the theoretical importance of the effect is not measured by its absolute size, but rather by the theoretical distinctions made over the course of the experimental investigation. Indeed, despite its size, the stimulus manipulation of Experiment 3 provided evidence that the role switch cost is attributable not only to differences at the *pose level* (i.e., switches in body posture), but also to a more abstract *relational level* (i.e., switches in event roles).

6.7. Conclusions

To close, over the course of four experiments, we have provided empirical evidence that the human visual system is spontaneously engaged in extracting the structure of what is happening in the world – including the interactive relationships between people. The rapidity of the extraction and its generality over a wide range of events suggests that this information may have a strong influence on how we describe the world and understand what we observe more generally.

Author Contributions
A. Hafri and B. Strickland developed the study concept. All authors contributed to the study design. A. Hafri performed testing, data collection, and data analysis. All authors contributed to interpretation of analyses. A. Hafri drafted the manuscript, and B. Strickland and J.C. Trueswell provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments
We thank the actors; Tim Dawson, Stamati Liapis, Estee Ellis, and Juliet Crain for assistance in

data collection; and Russell Epstein, Lila Gleitman, and the Trueswell and Epstein labs for helpful discussion.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was supported by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC (to B.S.); UPenn graduate research funds; and NSF Integrative Graduate Education and Research Traineeship, NSF Graduate Research Fellowship, and NIH Vision Training Grant 2T32EY007035-36 (to A.H.).

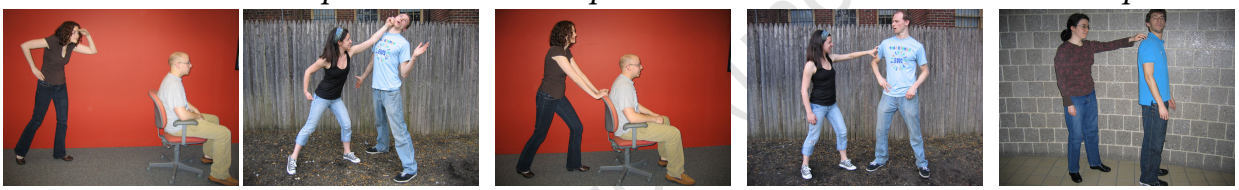
Appendix

An example image for each event category featured in the experiments (for Experiments 1a and 1b, Female Agent on the Left images; for Experiments 2 and 3, Blue Agent on the Left images). Agent and Patient poses were similar for the four versions of each event category. Although the images used in Experiments 2 and 3 were desaturated to a level of 3% to make the task (color search) more difficult, they are shown here in full color for illustrative purposes. See sections 2.1.2, 4.1.2, and 5.1.2 for details.

Experiments 1a and 1b (Gender Search)

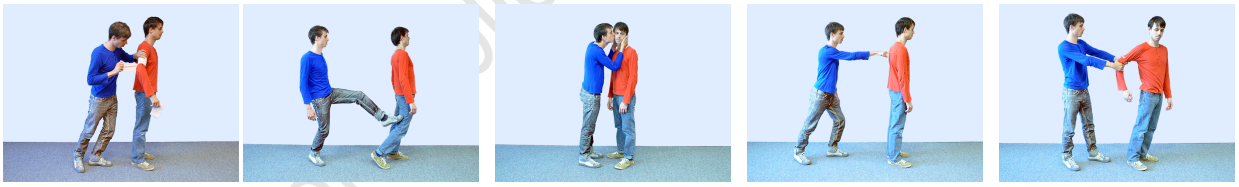
brush *chase* *feed* *film* *kick*

look at *punch* *push* *scratch* *tap*

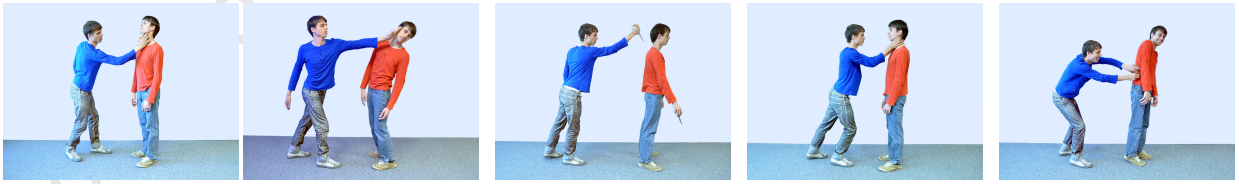


Experiment 2 (Color Search)

bandage *kick* *kiss* *poke* *pull*



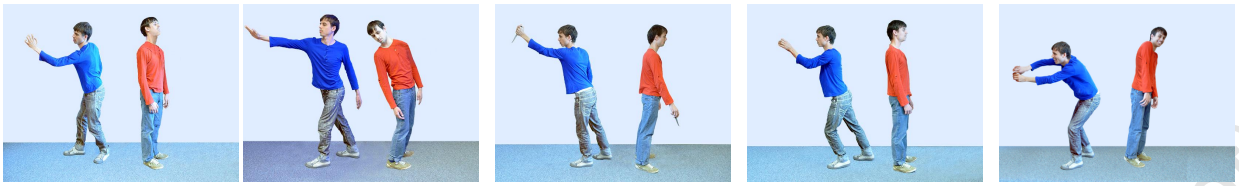
scratch *slap* *stab* *strangle* *tickle*



Experiment 3 (Color Search, Mirror-Flipped)

bandage *kick* *kiss* *poke* *pull*

scratch *slap* *stab* *strangle* *tickle*



References

Abrams, R. A., & Christ, S. E. (2003). Motion onset captures attention. *Psychological Science*, *14*(5), 427–432. http://doi.org/10.1111/1467-9280.01458

Aguirre, G. K. (2007). Continuous carry-over designs for fMRI. *NeuroImage*, *35*(4), 1480–94. http://doi.org/10.1016/j.neuroimage.2007.02.005

Baayen, R. H., & Milin, P. (2010). Analyzing Reaction Times. *International Journal of Psychology Research*, *3*, 12–28. http://doi.org/10.21500/20112084.80

Baillargeon, R., Stavans, M., Wu, D., Gertner, Y., Setoh, P., Kittredge, A. K., & Bernard, A. (2012). Object individuation and physical reasoning in infancy: An integrative account. *Language Learning and Development*, *8*(1), 4–46. http://doi.org/10.1080/15475441.2012.630610

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: the influence of trial history and data transformations. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(5), 1563–71. http://doi.org/10.1037/a0032186

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. http://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Green, P. (2016). lme4: Linear Mixed-Effects Models using "Eigen" and S4 (Version 1.1-12). Retrieved November 1, 2016, from https://cran.r-project.org/package=lme4

Biederman, I., Blickle, T. W., Teitelbaum, R. C., & Klatsky, G. J. (1988). Object search in nonscene displays. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 456–467. http://doi.org/10.1037/0278-7393.14.3.456

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–77.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436. http://doi.org/10.1163/156856897X00357

Brown, P. M., & Dell, G. S. (1987). Adapting Production to Comprehension: The Explicit Mention of Instruments. *Cognitive Psychology*, *19*, 441–472.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(4), 753–63. http://doi.org/10.1037/0096-1523.33.4.753

Cheries, E. W., Wynn, K., & Scholl, B. J. (2006). Interrupting infants' persisting object representations: An object-based limit? *Developmental Science*, *9*(5), 50–58. http://doi.org/10.1111/j.1467-7687.2006.00521.x

Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: the processing of semantic roles in visual narrative. *Cognitive Psychology*, *67*(3), 73–97. http://doi.org/10.1016/j.cogpsych.2013.07.002

Cohn, N., Paczynski, M., & Kutas, M. (2017). Not so secret agents: Event-related potentials to semantic roles in visual event comprehension. *Brain and Cognition*, *119*(April), 1–9. http://doi.org/10.1016/j.bandc.2017.09.001

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., … Haxby, J. V. (2012). The representation of biological classes in the human brain. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(8), 2608–18. http://doi.org/10.1523/JNEUROSCI.5547-11.2012

Croft, W. (2012). *Verbs: Aspect and Causal Structure*. Oxford: Oxford University Press.

Dobel, C., Diesendruck, G., & Bölte, J. (2007). How writing system and age influence spatial representations of actions: a developmental, cross-linguistic study. *Psychological Science*, *18*(6), 487–91. http://doi.org/10.1111/j.1467-9280.2007.01926.x

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619.

Dryer, M. S. (2013). Order of Subject, Object and Verb. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Fausey, C. M., Long, B. L., Inamori, A., & Boroditsky, L. (2010). Constructing agency: the role of language. *Frontiers in Psychology*, *1*(October), 162. http://doi.org/10.3389/fpsyg.2010.00162

Feldman, H., Goldin-Meadow, S., & Gleitman, L. R. (1978). Beyond Herodotus: The creation of language by linguistically deprived deaf children. In A. Lock (Ed.), *Action, Symbol, and Gesture: The Emergence of Language* (pp. 351–414). New York: Academic Press.

Fillmore, C. J. (1968). The Case for Case. *Texas Symposium on Language Universals*. http://doi.org/10.2307/326399

Fodor, J. A. (1983). *The Modularity of Mind*. Boston: MIT Press.

Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. *Perception & Psychophysics*, *33*(6), 575–581. http://doi.org/10.3758/BF03202940

Gervais, W. M., Reed, C. L., Beall, P. M., & Roberts, R. J. (2010). Implied body action directs spatial attention. *Attention, Perception & Psychophysics*, *72*(6), 1437–43. http://doi.org/10.3758/APP.72.6.1437

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, *4*(3), 179–92. http://doi.org/10.1038/nrn1057

Glanemann, R., Zwitserlood, P., Bölte, J., & Dobel, C. (2016). Rapid apprehension of the coherence of action scenes. *Psychonomic Bulletin & Review*. http://doi.org/10.3758/s13423-016-1004-y

Gleitman, L. R. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, *1*(1), 3–55. http://doi.org/10.1207/s15327817la0101_2

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(4), 544–569. http://doi.org/10.1016/j.jml.2007.01.007

Goldin-Meadow, S., So, W. C., Ozyürek, A., & Mylander, C. (2008). The natural order of events: how speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(27), 9163–8. http://doi.org/10.1073/pnas.0710060105

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1), 1–11. http://doi.org/10.1167/14.1.14.doi

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–76. http://doi.org/10.1016/j.cogpsych.2008.06.001

Gruber, J. S. (1965). Studies in lexical relations. In *PhD Dissertation*. Cambridge, MA.

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, *142*(3), 880–905. http://doi.org/10.1037/a0030045

Hafri, A., Trueswell, J. C., & Epstein, R. A. (2017). Neural representations of observed actions generalize across static and dynamic visual input. *The Journal of Neuroscience*, *37*(11), 2496–16. http://doi.org/10.1523/JNEUROSCI.2496-16.2017

Hafri, A., Trueswell, J. C., & Strickland, B. (2016). Extraction of event roles from visual scenes is rapid, automatic, and interacts with higher-level visual processing. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Philadelphia, PA.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–9. http://doi.org/10.1038/nature06288

Hartshorne, J. K. (2014). What is implicit causality? *Language, Cognition and Neuroscience*, *29*(7), 804–824. http://doi.org/10.1080/01690965.2013.796396

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the

posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 201714471. http://doi.org/10.1073/pnas.1714471114

Jackendoff, R. S. (1990). *Semantic Structures*. Cambridge, MA: MIT Press. http://doi.org/10.1037/031829

Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G., & Orban, G. a. (2010). Coding observed motor acts: different organizational principles in the parietal and premotor cortex of humans. *Journal of Neurophysiology*, *104*(1), 128–140. http://doi.org/10.1152/jn.00254.2010

Kako, E. (2006). Thematic role properties of subjects and objects. *Cognition*, *101*(1), 1–42. http://doi.org/10.1016/j.cognition.2005.08.002

Kline, M., Muentener, P., & Schulz, L. (2013). Transitive and periphrastic sentences affect memory for simple causal scenes, *1*, 1–5.

Kominsky, J. F., Strickland, B., Wertz, A. E., Elsner, C., Wynn, K., & Keil, F. C. (2017). Categories and Constraints in Causal Perception. *Psychological Science*. http://doi.org/10.1177/0956797617719930

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science*, *14*(5), 402–8.

Langacker, R. (1987). *Foundations of Cognitive Grammar*. Stanford: Stanford University Press.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *26*(11), 2894–906. http://doi.org/10.1523/JNEUROSCI.4915-05.2006

Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition*, *50*(1–3), 211–238. http://doi.org/10.1016/0010-0277(94)90029-9

Leslie, A. M. (1995). A Theory of Agency.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(April 1986), 265–288.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.

Levin, B., & Rappaport-Hovav, M. (2005). *Argument Realization*. Cambridge, UK: Cambridge University Press.

Mayrhofer, R., & Waldmann, M. R. (2014). Indicators of causal agency in physical interactions: The role of the prior context. *Cognition*, *132*(3), 485–490. http://doi.org/10.1016/j.cognition.2014.05.013

Mouchetant-Rostaing, Y., Giard, M.-H., Bentin, S., Aguera, P.-E., & Pernier, J. (2000). Neurophysiological correlates of face gender processing in humans. *European Journal of Neuroscience*, *12*(1), 303–310. http://doi.org/10.1046/j.1460-9568.2000.00888.x

Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. *Cognitive Psychology*, *61*(2), 63–86. http://doi.org/10.1016/j.cogpsych.2010.02.001

Nonyane, B. A. S., & Theobald, C. M. (2007). Design sequences for sensory studies: achieving balance for carry-over and position effects. *The British Journal of Mathematical and Statistical Psychology*, *60*(Pt 2), 339–349. http://doi.org/10.1348/000711006X114568

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. http://doi.org/10.1023/A:1011139631724

Oosterhof, N. N., Tipper, S. P., & Downing, P. E. (2012). Viewpoint (in)dependence of action representations: an MVPA study. *Journal of Cognitive Neuroscience*, *24*(4), 975–89. http://doi.org/10.1162/jocn_a_00195

Oosterwijk, S., Winkielman, P., Pecher, D., Zeelenberg, R., Rotteveel, M., & Fischer, A. H. (2012). Mental states inside out: switching costs for emotional and nonemotional sentences that differ in internal and external focus. *Memory & Cognition*, *40*(1), 93–100. http://doi.org/10.3758/s13421-011-0134-8

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, *108*(1), 155–84. http://doi.org/10.1016/j.cognition.2008.02.007

Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The Two-Body Inversion Effect. *Psychological Science*, 1–11. http://doi.org/10.1177/0956797616685769

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. *Psychological Science*, *14*(2), 119–124. http://doi.org/10.1111/1467-9280.t01-1-01429

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. http://doi.org/10.1163/156856897X00366

Pinker, S. (1989). Learnability and Cognition: The Acquisition of Argument Structure. In *Language* (Vol. 68, p. xiv, 411).

Potter, M. C. (1976). Short-Term Conceptual Memory for Pictures. *Journal of Experimental Psychology : Human Learning and Memory*, *2*(5), 509–522.

Rissman, L., Rawlins, K., & Landau, B. (2015). Using instruments to understand argument structure: Evidence for gradient representation. *Cognition*, *142*, 266–290. http://doi.org/10.1016/j.cognition.2015.05.015

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Reviews. Neuroscience*, *11*(4), 264–74. http://doi.org/10.1038/nrn2805

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology*, *23*(3), 250–254. http://doi.org/10.1016/j.cub.2012.12.017

Ross, H. (1972). *Play It Again, Sam*. United States: Paramount Pictures.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1–2), 1–46. http://doi.org/10.1016/S0010-0277(00)00152-9

Scholl, B. J., & Gao, T. (2013). Percieving animacy and intentionality: Visual processing or higher-level judgment? In M. D. Rutherford & V. A. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention*. MIT Press.

Scholl, B. J., & Leslie, A. M. (1999). Modularity, Development and "Theory of Mind." *Mind and Language*, *14*(1), 131–153. http://doi.org/10.1111/1468-0017.00106

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. http://doi.org/10.1037/0033-295X.84.2.127

Shirai, N., & Imura, T. (2016). Emergence of the ability to perceive dynamic events from still pictures in human infants. *Scientific Reports*, *6*(November), 37206. http://doi.org/10.1038/srep37206

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. http://doi.org/10.1111/j.1467-7687.2007.00569.x

Spence, C., Nicholls, M. E., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics*, *63*(2), 330–336. http://doi.org/10.3758/BF03194473

Strickland, B. (2016). Language reflects "core" cognition: A new theory about the origin of cross-linguistic regularities. *Cognitive Science*, 1–32. http://doi.org/10.1111/cogs.12332

Strickland, B., & Scholl, B. J. (2015). Visual Perception Involves Event-Type Representations: The Case of Containment Versus Occlusion. *Journal of Experimental Psychology: General*, *144*(3), 570–580.

Talmy, L. (2000). *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*. http://doi.org/10.1038/381520a0

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*(3), 209–253. http://doi.org/10.1016/S0010-0277(99)00069-4

Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language*, *63*(1), 64–82. http://doi.org/10.1016/j.jml.2010.02.006

Tucciarelli, R., Turella, L., Oosterhof, N. N., Weisz, N., & Lingnau, A. (2015). MEG multivariate analysis reveals early abstract action representations in the lateral occipitotemporal cortex. *Journal of Neuroscience*, *35*(49), 16034–16045. http://doi.org/10.1523/JNEUROSCI.1422-15.2015

van Buren, B., Uddenberg, S., & Scholl, B. J. (2015). The automaticity of perceiving animacy: Goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychonomic Bulletin & Review*. http://doi.org/10.3758/s13423-015-0966-5

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454–61.

Verfaillie, K., & Daems, A. (1996). The priority of the agent in visual event perception: On the cognitive basis of grammatical agent-patient asymmetries. *Cognitive Linguistics*, *7*(1996), 131–148. http://doi.org/10.1515/cogl.1996.7.2.131

White, A. S., Reisinger, D., Rudinger, R., Rawlins, K., & Durme, B. Van. (2017). Computational linking theory.

Wilson, F., Papafragou, A., Bunger, A., & Trueswell, J. (2011). Rapid extraction of event participants in caused motion events. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX.

Wurm, M. F., & Lingnau, A. (2015). Decoding actions at different levels of abstraction. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(20), 7727–7735. http://doi.org/10.1523/JNEUROSCI.0188-15.

Yin, J., & Csibra, G. (2015). Concept-Based Word Learning in Human Infants. *Psychological Science*, *26*(8), 1316–24. http://doi.org/10.1177/0956797615588753

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology. General*, *138*(2), 307–327. http://doi.org/10.1037/a0015305

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–93. http://doi.org/10.1037/0033-2909.133.2.273

Zheng, M., & Goldin-Meadow, S. (2002). Thought before language: How deaf and hearing children express motion events across cultures. *Cognition*, *85*(2), 145–175. http://doi.org/10.1016/S0010-0277(02)00105-1