Implicit social associations for geometric-shape agents more strongly influenced by visual form than by explicitly identified social actions

Ben Kenward[1,2*], Mathias Berggren[2], Michiteru Kitazaki[3], Shoji Itakura[4], Yasuhiro Kanakogi[5]

In press at Psychologia (Accepted February 7)

[1]Department of Psychology, Oxford Brookes University, UK

[2]Department of Psychology, Uppsala University, Sweden

[3]Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

[4]Graduate School of Letters, Kyoto University, Japan

[5]Department of Psychology, Otemon Gakuin University, Japan

*Correspondence:
Ben Kenward
Department of Psychology
Oxford Brookes University
Headington Campus
Oxford
OX3 0BP
UK
ben.kenward@wolfson.oxon.org.

Acknowledgements

Thanks to Ida Kenward for assistance with data collection, and to Nazar Akrami for comments on the manuscript.

ABSTRACT

Studies of infants' and adults' social cognition frequently use geometric-shape agents such as coloured squares and circles, but the influence of agent visual-form on social cognition has been little investigated. Here, although adults gave accurate explicit descriptions of interactions between geometric-shape aggressors and victims, implicit association tests for dominance and valence did not detect tendencies to encode the shapes' social attributes on an implicit level. With regard to valence, the lack of any systematic implicit associations precludes conclusive interpretations. With regard to dominance, participants implicitly associated a yellow square as more dominant than a blue circle, even when the true relationship was the reverse of this and was correctly explicitly described by participants. Therefore, although explicit dominance judgements were strongly influenced by observed behaviour, implicit dominance associations were more clearly influenced by preconceived associations between visual form and social characteristics. This study represents a cautionary tale for those conducting experiments using geometric-shape agents.

Keywords: Geometric-shape agents, visual form, social judgement, intention attribution, implicit association test

INTRODUCTION

Animated geometric-shape agents with no morphology in common with humans or other animals are readily interpreted as intentional social agents if they move like social agents do (Abell et al., 2000; Barrett et al., 2005; Gao et al., 2009; Heberlein and Adolphs, 2004; Scholl and Tremoulet, 2000). In one classic example a featureless triangle was described as “dominating, power-loving, [and] possessive” due to its self-propelled movements in interaction with other geometric shapes (Heider and Simmel, 1944). From birth, such biological motion is processed differently to mechanical motion, prompting the conclusion that evolved specialised mechanisms form the basis for such processing (Frankenhuis and Barrett, 2013; Gelman, 1990; Simion et al., 2013).

Because adults readily interpret moving geometric shapes as agents, such shapes have been used in a number of recent studies of adults’ social cognition in which geometric-shape processing in itself was not the focus, simply because geometric shapes are convenient for producing uncomplicated social stimuli. For example, geometric-shape agents have been used to investigate cultural differences in theory of mind (Koelkebeck et al., 2011) and processing of norm violations and punishment (Eriksson et al., 2016). Geometric shape agents have also been used to study the perception of social events like chasing (Gao et al., 2010; Gao et al., 2009) and the roles of spatial and temporal cues in judgements of intentionality (Scholl and Tremoulet, 2000; Takahashi and Watanabe, 2015). However adults’ affective valence responses and other implicit responses to geometric shapes have been less studied (Kuhlmeier, 2013). It is relevant to note that manipulation of human stimuli to make them look less human diminishes affective valence responses (Mould et al., 2012). It is therefore plausible that although adults’ explicit reports concerning geometric-shape agents are similar to their reports for real agents, their implicit associations may be divergent. This might present a methodological concern for the use of geometric-shape agents. This study therefore investigates adults’ implicit valence associations and implicit social intention ascriptions for geometric-shape agents, and compares these implicit responses to explicit reports.

Because infants also readily interpret moving geometric shapes as agents, and because such simple stimuli can be easily controlled with respect to their superficial perceptual properties, geometric shapes have also been widely used in studies of infants' social cognition. Not only have they been used to study basic questions of infants' understanding of animacy and goal-directed movement (Biro et al., 2007; Gergely and Csibra, 2003; Luo et al., 2009; Premack, 1990), but they have also been widely used to investigate infants' judgements and expectations of agents' social actions, including value judgements (e.g. Kuhlmeier et al., 2003; Lyons and Cheries, 2017; Meristo and Surian, 2013; Premack and Premack, 1997; Tatone et al., 2015; Thomsen et al., 2011). In one example of such a study, six- and ten-month-olds viewed two geometric-shape agents who either helped or hindered a third such agent in achieving its goal of climbing a hill (Gredebäck et al., 2015; Hamlin, 2015; Hamlin et al., 2007). Given the choice of reaching for the agents, infants preferred the helper over the hinderer and even over a neutral agent. Infants themselves help a geometric-shape agent who is unable to reach its goal (Kenward and Gredeback, 2013). Such studies are taken as indicating that infants have a preference for agents who behave prosocially (Hamlin, 2013; Kanakogi et al., 2017).

In a further such study, ten-month-olds viewed two interacting geometric-shape agents, one of whom behaved aggressively toward the other by chasing, knocking, and squashing it (Kanakogi et al., 2013). Given the choice of reaching for the agents, infants preferred the victim over the aggressor, and even preferred the victim over a neutral agent, meaning the results cannot be explained only by a negative evaluation of an antisocial act. The authors concluded that this preference for a victim is best explained by a rudimentary form of sympathy for victims of antisocial acts, because empathic detection of others' suffering leads to sympathetic approach behaviour (de Waal, 2008; 2012; Fujisawa et al., 2006; Kanakogi et al., 2017).

Studies of infants have therefore established not only that they interpret the actions of geometric shapes in terms of goals, but also that they evaluate these goals in terms of valence, leading to approach or avoidance. When it comes to adults, however, we are not aware of a published study which examines approach or avoidance of geometric-shape agents, although there is at least one study demonstrating adults can make explicit value judgements concerning such shapes (Eriksson et al., 2016).

A priori study goals

The original goal of this study was to investigate adults' attributions of behavioural intention and valence to geometric-shape agents, the associations between attributed intention and valence, and the associations between implicit and explicit attributions. It is currently unclear how closely coupled intention attribution mechanisms are to valence attribution mechanisms in situations in which observers are aware of the unreality of the situation. While it is obviously the case that fictional drama can elicit powerful positive or negative emotional reactions, it is also plausible that unreal social scenarios such as those featuring geometric shapes may be more effective at eliciting intention attribution than explicit preference judgements. Affective responses need not arise from the understanding of the consequences of social actions – these processes are independent and dissociable, both behaviourally and neurologically (Gonzalez-Liencres et al., 2013; Shamay-Tsoory et al., 2009). In fact, a pilot study (Kanakogi et al., 2012) which tested adults' reactions to stimuli which in infants had elicited preference for a victim and avoidance of an aggressor (Kanakogi et al., 2013) suggested that adults did not share this spontaneous preference, although they explicitly attributed relevant intentions to the interacting shapes. Possibly, explicit preferences in adults were lacking because of the explicit knowledge of the unreality of the situation. However, explicit knowledge of unreality might not suppress any implicit preferences which may be present.

We utilised stimuli developed to test infants' social preferences between a victim and a dominant aggressor, because these stimuli are already known to generate preferences (in infants, Kanakogi et al., 2013) and explicit attributions of intentionality (in adults, Kanakogi et al., 2012). We followed the pilot study (Kanakogi et al., 2012) in testing explicit preference in a way as close as possible to the original infant study, and by testing explicit intention attribution by simply asking what participants thought of the shapes. We also asked participants to explicitly justify their preferences. For implicit associations, we used two implicit association tests (IAT) (Greenwald et al., 1998). One test uses a valence (good / bad) concept pair for testing implicit preference, and one uses a dominance / subordinance concept pair for testing behavioural intention attribution (because explicit responses in the pilot study frequently fell into this category).

Implicit judgement of geometric agents

The IAT has shown good reliability and correlation with explicit scales and behaviours in many contexts (De Houwer et al., 2009; Greenwald et al., 2009; Nosek et al., 2007), although the issue of implicit/explicit attitude correlation is not without controversy (Oswald et al., 2013). It measures the relative associations of two target concepts (here the victim and the aggressor) with two attribute concepts (here good and bad or dominant and subordinant in the two IATs). It utilises the fact that reaction-times for quickly sorting different exemplars using the same key-press are faster when the exemplars are more closely implicitly associated. Here, the IAT could reveal for example that the victim is associated as more positive than the aggressor if participants react more quickly when sorting victims and positive words using the same computer key (congruent condition) than when sorting aggressors and positive words using the same key (incongruent condition). While the valence IAT is intended simply to explore potential implicit valence associations (see below for predictions), the dominance IAT was further intended as a manipulation check concerning the stimuli. If adult participants view the agent interactions as clearly reflecting aggressive dominance, as they are intended to and as pilot data suggested, then the dominant aggressor should be implicitly associated with dominance.

In line with the pilot study (Kanakogi et al., 2012), we predicted that participants would implicitly and explicitly correctly attribute aggressive dominance to geometric-shape agents, but that this would not necessarily lead to systematic explicit preference against such agents, due to knowledge of the unreality of the situation (in contrast to infants, Kanakogi et al., 2013). However, according to our hypothesis that this knowledge might not influence implicit valence judgements, we predicted that the implicit measures might reveal preference for the victim over the aggressor. This prediction is in line with the idea that infant behaviour sometimes reflects implicit processes which develop early but show continuity into adulthood because they are highly functional (Spelke and Kinzler, 2007). Further, although we do not predict strong systematic explicit preferences, because implicit preferences can unconsciously influence explicit preferences (Greenwald et al., 2009; Nosek et al., 2007), we predicted that there might be a correlation between the two.

Further, a number of personality traits would be expected to modulate preference for a victim over an aggressor, so finding relations between these traits and implicit or explicit preference would provide validation of these preference measures. We hypothesised that lower social dominance orientation (Pratto et al., 1994), greater empathy, and lower aggression might correlate with increased preference for the victim, and we therefore measured these by questionnaire.

Post-hoc study goals

Although we counter-balanced the forms of the victim and aggressor (blue circle versus yellow square), exploratory data examination indicated that visual form had strong effects on IAT results. Although this was not originally intended to be the study focus, we therefore also systematically examined whether visual form affects implicit and explicit preference and intention attribution. Previous studies have demonstrated valence associations with specific (non-agent) abstract shapes, with more curved and less angular objects having more positive valence (Bar and Neta, 2006; Larson et al., 2012). However, although those results suggest that visual form could be important for the processing of geometric-shape agents, we are not aware of any study which has focussed on this issue. Although this issue was not the original focus of this study, because of unpredicted results indicating that effects of visual form were important in our data, discussion focuses on this issue.

METHOD

Participants

One hundred and thirty eight students (57 men; M_{age} = 24, SD_{age} = 5) participated, with 25 Japanese participants recruited in Kyoto (part of the unpublished pilot study, Kanakogi et al., 2012), 21 Swedes in Stockholm and 92 Swedes in Uppsala. Of these participants, all were assessed using explicit measures, but a subset of 64 (all from Uppsala; 23 men; M_{age} = 24, SD_{age} = 5) completed an IAT and personality questionnaire, with half randomly assigned to complete the Dominance IAT and the other half the Valence IAT. The data were collected in two waves, with Japanese data and Swedish data without IAT collected before the IAT was designed. No participants were excluded from analysis. Sample sizes (and therefore data collection stopping rules) were determined by practical considerations, chiefly the amount of data collectable within the course of a Swedish undergraduate dissertation project. Local and international (Declaration of Helsinki, American Psychological Association) ethical guidelines were followed, including approval by the Uppsala Regional Ethical Review Committee, and participants gave written informed consent.

Materials

Familiarisation stimuli showed interactions between a blue ball and a yellow cube within a static green frame displayed on computer. Familiarisation sequences consisted of two alternating 20 second clips, in which the aggressor chased, knocked, and finally crushed the victim against the frame (the two clips differed only in the exact movement paths taken by the agents). Participants not taking the IAT and questionnaire saw the exact same familiarisation sequence as previously displayed to infants, with three of each clip variant presented for a total of six clip presentations, and brief attention grabbers (rotating cartoon animal faces with sound effects) between each clip (see Kanakogi et al., 2013 for further details of familiarisation stimuli - a video clip can be viewed at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0065292#s6). The familiarisation sequences used in the IAT were composed of the same two clips, with two of each variant for a total of four, without attention grabbers. Fewer clips were used per sequence because the sequence was displayed more than once (see below). Physical replicas of the agents were used for the preference trials, which were the same objects as used previously (Kanakogi et al., 2013), measuring 6.5cm across. Aggressor and victim identity were counterbalanced (blue circle or yellow square; shape and colour were not separately counterbalanced).

The IAT was administered in the laboratory using OpenSesame 0.27.4, which has good performance with respect to latency (Mathôt et al., 2012). Stimuli were sorted with the Z and M computer keys. The IAT target stimuli consisted of pictures of the ball or cube within the green frame, with four variants of each with the agents in different locations in the frame. The attribute stimuli were Swedish words (translated here), positioned centrally, with eight words for each attribute. Valence words were from standard IAT lists, but with words specific to social interaction, such as Love, replaced; good – Wonderful, Happiness, Good, Pleasure, Paradise, Fantastic, Excellent, Pleasant; bad – Horrible, Terrible, Unpleasant, Death, Disease, Catastrophe, Awful, Bad. Dominance and subordinance word lists were constructed for the present task; dominance – Power, Strength, Ruler, Victory, Dominance, Advantage, Boss, Superior; subordinance – Surrender, Weakness, Defeat, Victim, Loser, Slave, Oppressed, Inferior.

The IAT operated as follows: target sorting practice 1 (agents only, 32 trials); test block 1 (64 trials, 32 each for agents and attribute words); target sorting practice 2 (agents only, 32 trials, sorting keys reversed from practice 1); test 2 (64 trials, 32 each of agents and attribute words, agent sorting keys still reversed but attribute sorting keys not reversed).

Before every practice and test block a four-clip familiarisation sequence was displayed. Before test block 1 participants had therefore seen 8 clips establishing the agent’s roles (160 seconds) and before test block 2 participants had seen 16 clips (320 seconds). Although a number of studies have demonstrated that the IAT can measure associations created during an experimental session (De Houwer et al., 2007; Han et al., 2010; Han et al., 2006), the IAT has most commonly been used to test associations that existed prior to the experimental session and are therefore possibly stronger. These interspersed familiarisation sequences were therefore intended to maximise the strength of participants’ implicit associations for the agents during the test. This is also why we omit training on the attribute word sorting, which could have diluted the salience of the agents’ roles. The IAT is standard apart from these modifications.

The questionnaire consisted of 49 Likert items from three instruments: the full Social Dominance Orientation instrument (Pratto et al., 1994); the Aggression Questionnaire – Revised Short Version (Ågren and Prochazka, 2001; Bryant and Smith, 2001); and the subscales Empathic Concern, Perspective Taking, and Fantasy from the Interpersonal Reactivity Index (Cliffordson, 2002; Davis, 1996).

Procedure

Participants not taking the IAT and questionnaire were shown the familiarisation sequence and then immediately shown the two physical replica objects, with the experimenter saying "Which do you choose? Go on gut-feeling". (This wording was chosen because the original pilot study, which comprised the Japanese sample, aimed to maximise comparability between infant and adult methods, although this was not important here.) The choice was immediately noted, and after the choice, the experimenter asked the participant on what basis they made their choice, and then asked them to describe their impression of each agent and the relationship between them, writing down the answers verbatim. Participants taking the IAT and questionnaire underwent exactly the same procedure immediately after taking the IAT, except that two instead of six clips were shown in the familiarisation sequence immediately prior to choice because the clips had been displayed so many times during the IAT.

The following variables were counterbalanced: aggressor/victim identity, initial agent and IAT sorting key pairings, IAT block order (congruent/incongruent), agent presentation side in physical choice, and whether the questionnaire was completed before or after the IAT and physical choice. Trial order within blocks was randomised.

Coding and analysis

Agent choice justifications were coded from written notes according to the following categories: (1) strong, active, dominant; (2) weak, gentle, nice, pitiable; (3) aesthetic, (4) other, or (5) don't know. The first two categories were scored if the stated words or close synonyms were used; aesthetic was scored if the agents' shape, colour, or texture was referred to; other was scored for any other reason except don't know. Although not a necessary consequence of this scheme, it was the case that all explicitly social justifications fell into categories 1 or 2, not 4 (other).

Agent and interaction descriptions were scored from notes for whether either were ascribed the following three increasingly specific (mutually inclusive) properties: social (any explicitly social language), socially asymmetric (teasing/teased, strong/weak, hunter/hunted, or synonyms), and dominance related (strong/weak, hunter/hunted, or synonyms).

For each individual an IAT *D* score was calculated according to the improved standard recommended algorithm (Greenwald et al., 2003), the key steps of which are subtracting the mean reaction time in the congruent block (valence IAT: victim paired with positive; dominance IAT: victim paired with subordinance) from the mean reaction time for the incongruent block, and dividing by the pooled standard deviation, with the reaction times for incorrect trials replaced with 600 ms plus the mean for that block.

Our primary predictions (see introduction) were that explicit and implicit measures would or would not reveal group-level tendencies to identify the agents as more or less dominant or positively valenced. Our primary analyses are therefore comprised of descriptive estimates (with confidence intervals) of group tendencies to explicitly or implicitly identify agents as more or less dominant or positively valenced.

RESULTS

Preliminary analyses found no effects of age or culture on any dependent variables (details omitted for brevity, all p-values > .1).

Explicit agent choice, choice justification, and agent description

Overall, there was no significant bias towards choosing either agent, with 54% of participants, 95% CI [46%, 62%], choosing the victim, $p = .349$, sign test. A significant majority of participants chose the blue sphere rather than the yellow cube, 59%, 95% CI [51%, 67%], $p = .033$, sign test. The agents or their interaction were described as social by 99% of participants, as socially asymmetric by 96% of participants, and as dominance related by 86% of participants. No participant gave a description of the agents or their actions in which the roles were incorrectly reversed. Justifications for choice of agent are presented in Table 1. Personality variables had the expected pattern of correlations (for example between aggression and social dominance orientation), establishing some validity for these measures in our sample, but showed no correlations with agent choice (Table 2).

Table 1.

Agent choice justifications (percentage participants), by agent choice.

Agent choice	Strong, active, dominant	Weak, gentle, nice, pitiable	Aesthetic	Other	Don't know
Aggressor	57	0	40	30	2
Victim	7	41	39	32	3

Row totals are greater than 100% because more than one justification was possible.

Table 2.

Correlation coefficients for personality, agent choice, agent choice justification, and IAT D-score variables

	1	2	3	4	5	6	7	8	9	10	11
1. Social dominance orientation											
2. Aggression	.47***										
3. Fantasy	-.02	.23									
4. Empathic concern	-.44***	-.18	.15								
5. Perspective taking	-.11	-.18	.16	.33**							
6. Explicit agent choice: victim	-.12	.06	.13	-.07	-.17						
7. Agent choice justification: strong	.15	.08	.07	-.15	.10	-.55***					
8. Agent choice justification: weak	-.14	-.08	.16	.16	-.04	.49***	-.35***				
9. Agent choice justification: aesthetic	-.06	-.05	-.08	-.03	.21	-.01	-.26**	-.40			
10. IAT *D*: dominance (aggressor)	-.05	.03	.00	.31	.00	.16	-.15	.25	-.06		
11. IAT *D*: dominance (yellow square)	.30	-.03	-.08	-.03	-.30	.27	-.41*	.10	.12	.28	
12. IAT *D*: valence	.06	.04	-.02	-.19	-.13	.21	-.27	.09	-.05	-	-

Notes. * $p < .05$; ** $p < .01$; *** $p < .001$. Correlations are Pearson's *r* except for those involving variables 6, 7, 8, and 9 which are point biserial correlations as these variables are binary. $N = 138$, except $N = 64$ for those involving personality variables (1 to 5) and $N = 32$ for those involving IAT *D* variables (10 to 12). IAT *D* dominance (yellow square) is the post-hoc variable showing the extent to which the yellow square is associated with dominance (see main text).

Implicit associations between the agents and concepts of dominance and valence

In the dominance IAT, a positive *D* score indicates faster reactions when the aggressor rather than the victim was paired with dominance words, indicating a stronger association between the aggressor and the concept of dominance. There was no group level tendency towards a positive *D* score, *M* = .04, *SD* = .42, which is not significantly greater than 0, *t*(31) = .64, *p* = .527, Cohen's *d* = .09, 95% CI for *d* [-.27, .45]. However, dominance *D* was strongly affected by the identity of the aggressor, with the yellow cube more strongly associated with dominance, irrespective of which agent was the actual aggressor (Fig. 1). This was confirmed by reversing the sign of the dominance *D* score for participants for whom the blue sphere was the aggressor, yielding a *D* for which positive scores meant faster reactions when the yellow cube was paired with dominance words. This *D* score was significantly positive, *M* = .26, *SD* = .33, *t*(31) = 4.55, *p* < .001, Cohen's *d* = .80, 95% CI for *d* [.44, 1.17]. The majority of participants, 75%, 95% CI [57%, 89%], had a *D* score consistent with associating the yellow cube more strongly with dominance, *p* = .007, sign test.

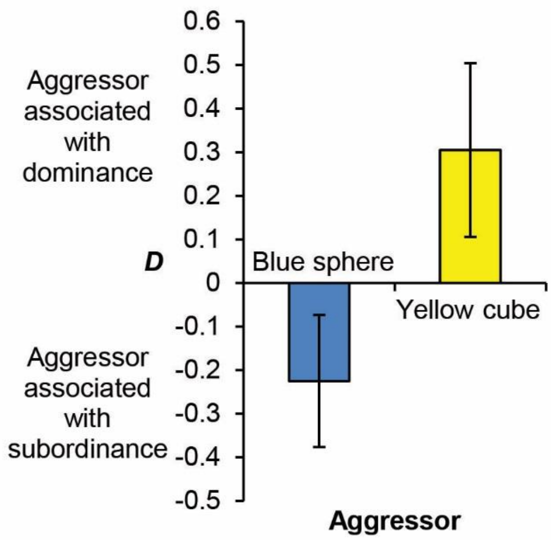

Fig. 1. Dominance IAT *D*-scores, by aggressor identity. Positive D scores imply an association between the aggressor and the concept of dominance. Error bars show 95% CIs.

Implicit judgement of geometric agents

In the valence IAT, a positive *D* score indicates faster reactions when the victim rather than the aggressor was paired with positive words, indicating a tendency to implicitly associate the victim as more positive. There was no significant group level tendency towards a positive *D* score, $M = .04$, $SD = .52$, which is not significantly greater than 0, $t(31) = .38$, $p = .705$, Cohen's $d = .07$, 95% CI for *d* [-.29, .43]. To investigate whether the explicitly preferred agent was implicitly associated as more positive, the sign of the valence *D* was reversed for those who chose the aggressor, yielding a *D* for which positive scores meant faster reactions when the chosen agent was paired with positive words. There was no significant group level tendency for this *D* to be positive, $M = .11$, $SD = .51$, $t(31) = 1.23$, $p = .230$, $d = .22$, 95% CI for *d* [-.14, .58]. To investigate whether the blue sphere was implicitly associated as more positive, the sign of the valence *D* was reversed for participants for whom the blue sphere was the aggressor. There was no significant group level tendency for this *D* to be positive, $M = .12$, $SD = .51$, $t(31) = 1.34$, $p = .191$, $d = .24$, 95% CI for *d* [-.12, .60].

Although *D* scores showed almost no significant correlations with other variables (Table 2), the *D* score indicating a tendency to associate the yellow square with dominance correlated negatively with the tendency to justify agent choice by describing an agent as strong, active, or dominant. This result was confirmed by comparing the relevant *D* scores between those who did ($M = .35$, $SD = .34$) or did not ($M = .05$, $SD = .18$) invoke this justification, $t(27) = 3.24$, $p = .003$.

DISCUSSION

In line with previous studies, the vast majority of adults here attributed social traits to geometric shapes, describing their movements with reference to human-like motives. Almost all participants had an accurate explicit memory of the social actions of the agents (and descriptions did not vary with personality or culture). Despite this, the dominance IAT did not appear to reflect this knowledge. No effect was detected, although the effect size confidence interval (upper bound .45) indicates the plausibility of a small undetected effect. On the other hand, three-quarters of participants had IAT scores indicating that they implicitly associated a yellow cube as more dominant than a blue sphere, irrespective of their explicit description of the actual relationship. We had not predicted such a strong result, but the effect size confidence interval (lower bound

.44) indicates that the effect is not spurious. Despite the modest sample size of 32, this confidence interval and the associated *p*-value ($p < .001$) strongly indicate the effect would prove reliable. Furthermore, even if an undetected effect of agent social behaviour does exist, the confidence intervals indicate that this undetected effect is smaller than the effect of agent visual form. The most important part of our discussion therefore focuses on this unexpected effect (below) but first we discuss our valence results in more detail.

There was no detected relation between participants' explicit agent choices and their valence IAT scores (although effect size confidence intervals indicate the plausibility of a medium undetected effect). Furthermore, for the valence IAT there were no detected systematic associations at all – not for behaviour or for visual form. Explicit preference between aggressor and victim also showed no systematic pattern. Roughly half the participants gave preference justifications appropriately referencing the agents' social intentions. However, we note that aesthetic justifications were also common, and that choice justifications are frequently post-hoc rationalisations which do not bear any relation to the true motivation for the choice (Hall et al., 2012; Johansson et al., 2005). We also note that the only systematic preference was based on visual form – the blue ball was weakly but significantly preferred, which is consistent with reports of a general preference for blue (Hurlbert and Ling, 2007; Palmer and Schloss, 2010). Together, these data do not provide strong evidence that intention attribution was clearly motivating implicit or explicit preference. A further observation is relevant: if choices were truly related to the agents' ascribed intentions, one might expect choice to be predicted by the traits empathy, aggression, or social dominance orientation, whereas post-hoc rationalisation predicts no such relations. No such relations were observed.

We note that there might be methodological reasons why the valence IAT as implemented was not optimal. The approach was motivated by the well-known association between approach and positive valence, together with an assumption that sympathy for victims is associated with approach. These associations do not, however, necessarily imply a simple link between sympathy for victims and positive valence – sympathy for a victim is also associated with negative feelings. We refrain from further discussion of the valence results as these issues together with the lack of systematic implicit or explicit preference present a challenge to clear interpretation.

With respect to valence, we were therefore unsuccessful in achieving our a priori study goal of comparing implicit and explicit responses. However, the unexpected but very clear effect of agent visual form on implicit dominance association means post-hoc discussion of this issue is warranted. As far as we know abstract shapes have not previously been found to be associated with specific social traits due to their visual form. It has previously been demonstrated that objects with sharper corners can be more aversive, however, and it has been argued that this is because such shapes are experienced as more threatening (Bar and Neta, 2006; Larson et al., 2012). This argument fits well with and receives support from the current IAT results and with the finding of a preference for the ball in the choice test. Further, given that in both the natural and the artificial worlds yellow is associated with danger (Stevens and Ruxton, 2012), it is plausible that colour may have contributed to the effect. There is therefore evidence lending plausibility to putative effects of both colour and shape in generating implicit dominance attribution, and it is not possible to establish their relative influence – we suggest that the strength of the effect may be due to a combination of both factors. Although it is possible that some similarity not directly related to dominance caused the association between visual form and the dominance words (IAT recoding; Meissner and Rothermund, 2013), other similarities are lacking in this case, rendering a non-dominance-based account implausible. Valence might have been a candidate, but there were no detected valence associations.

Although lacking precedent with respect to geometric-shape agents, it is in hindsight not puzzling that shape and/or colour produced dominance associations, given that shapes and colours can have socially relevant associations (Bar and Neta, 2006; Larson et al., 2012; Stevens and Ruxton, 2012). Importantly, we also note that several previous studies have shown that even infants form expectations about geometric-shape agents based on their visual forms (Lyons and Cheries, 2017; Thomsen et al., 2011). What does require explanation, however, is why objects should so strongly elicit such implicit associations in adults despite explicit descriptions of the agents entirely at odds with these associations.

Implicit judgement of geometric agents

The standard IAT method has been demonstrated in many different contexts to demonstrate a good fit between implicit and explicit attitudes (Greenwald et al., 2009; Nosek et al., 2007), although this issue is not uncontroversial (Oswald et al., 2013). This study joins some others in demonstrating less straightforward results (De Houwer et al., 2009) – factors which have no bearing on participant's explicit attitudes are known to sometimes influence IAT results (Han et al., 2006; Olson and Fazio, 2004). Previously, however, when factors unrelated to personal explicit attitudes have influenced implicit attitudes, the strongest influence on IAT results was nevertheless personal explicit judgement (Han et al., 2006). Here, however, participants' explicit attributions of dominance to the objects had no detectable effects on their dominance IAT performance, and any undetected effects can confidently be concluded to be weaker than the strong implicit associations between abstract visual form and dominance. This apparent disconnection between explicit and implicit associations requires further work to explain, and we now raise issues which should be addressed in future research.

Firstly, we consider a methodological issue: the majority of previous IAT studies measure attitudes which are well-established and therefore well-consolidated into memory (Greenwald et al., 2009). It is possible that here, the comparatively recently formed intention representations were not yet consolidated into automatic associations capable of reliably influencing IAT performance (Gawronski and Bodenhausen, 2006; Han et al., 2006). While this explanation cannot be ruled out, we do not favour it, because a number of previous studies, some using very similar methods, have indicated that attitudes created in the experimental session do influence IAT performance (De Houwer et al., 2007; Han et al., 2010; Han et al., 2006). A second methodological issue is that some aspects of the IAT method used here was slightly non-standard. It is possible that these modifications may have diminished the sensitivity of the IAT to detect any associations. However, this possibility appears very unlikely given that the dominance IAT did clearly detect associations with agent visual form.

Given that the reason that explicit intention representations did not influence IAT score is arguably not methodological, it is plausible that the apparent explicit / implicit disconnection reflects something real about intention attribution to geometric-shape agents. We suggest that although biological intention detection mechanisms readily process agents that observers know are in fact not really agents, the knowledge of the unreality may cause the detected intentions not to be encoded more deeply, in a way which could affect implicit associations. The suggestion that the social actions of geometric-shape agents do not lead to deeply encoded intention representations is consistent with the results of some studies of reactions to human social scenes. In these studies images are manipulated (e.g. by slight blurring) so that although the scene is recognisable, the subjects no longer appear realistically human. Participants report less strongly valenced emotional responses to such stimuli (Mould et al., 2012), and brain activation in regions associated with social and emotional processing is diminished (Mar et al., 2007).

Although implicit dominance associations were not related to the way in which participants described the agents when specifically questioned about their behaviour, when participants were asked for a more open-ended response justifying their choice of agent, a curious relation with the dominance IAT was found. Those who more strongly associated the yellow square with dominance were less likely to invoke an agent description of strong, active, or dominant as justifying their choice. Although the specific mechanism of this effect cannot yet be determined, the result suggests that implicitly activated dominance concepts may have interfered somehow with the tendency to explicitly invoke dominance. This result suggests that although the implicit dominance associations based on visual form cannot over-ride explicit knowledge that they conflict with, they can nevertheless affect spontaneous behaviour in more subtle ways. We note also that this correlation between implicit and explicit behaviour provides further evidence that implicit association of visual form with dominance is not an irrelevant artefact.

We now summarise our conclusions and their implications. We draw no strong conclusions with regards to valence, because there were few clear systematic group-level results for implicit or explicit measures. With regard to dominance, however, we found that although adults explicitly accurately described the social relationship between the agents, their implicit associations more strongly reflected apparently preconceived notions connecting visual form to social attributes. We cannot completely rule out methodological explanations for why implicit associations relevant to the shapes' intentions were not detected, although we view this as unlikely.

We finally note that these conclusions are specific to the current stimulus set, and may not generalise to other stimuli. We acknowledge that adults can make value judgements about unreal agents, whether in the form of geometric shapes (Eriksson et al., 2016), schematic cartoons (Buon et al., 2013), or described hypothetically (Bruers and Braeckman, 2014). To our knowledge, this study does not raise validity issues for specific previous studies using geometric-shape agents. However, we do suggest that the current results imply that studies using geometric-shape agents should be treated with a degree of caution pending further study, especially if adult implicit measures are to be used. Moral judgements of situations known to be unreal do differ from those concerning equivalent real situations (FeldmanHall et al., 2012). The disconnection between explicit and implicit associations suggested here could be one reason for this. If nothing else, the clear result that implicit associations between geometric-shape agents and social traits are strongly influenced by their visual form should be taken into account in the design of future studies.

REFERENCES

Abell, F., Happé, F., and Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development* 15**,** 1-16.

Ågren, H., and Prochazka, H. (2001). Aggression in the general Swedish population, measured with a new self-rating inventory: The Aggression Questionnaire - revised Swedish version (AQ-RSV). *Nordic Journal of Psychiatry* 55**,** 17-23.

Bar, M., and Neta, M. (2006). Humans Prefer Curved Visual Objects. *Psychological Science* 17**,** 645-648.

Barrett, H.C., Todd, P.M., Miller, G.F., and Blythe, P.W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior* 26**,** 313-331.

Biro, S., Csibra, G., and Gergely, G. (2007). "The role of behavioral cues in understanding goal-directed actions in infancy," in *From Action to Cognition,* eds. C. Von Hofsten & K. Rosander.), 303-322.

Bruers, S., and Braeckman, J. (2014). A Review and Systematization of the Trolley Problem. *Philosophia* 42**,** 251-269.

Bryant, F.B., and Smith, B.D. (2001). Refining the Architecture of Aggression: A Measurement Model for the Buss–Perry Aggression Questionnaire. *Journal of Research in Personality* 35**,** 138-167.

Buon, M., Jacob, P., Loissel, E., and Dupoux, E. (2013). A non-mentalistic cause-based heuristic in human social evaluations. *Cognition* 126**,** 149-155.

Cliffordson, C. (2002). The hierarchical structure of empathy: Dimensional organization and relations to social functioning. *Scandinavian Journal of Psychology* 43**,** 49-59.

Davis, M. (1996). *Empathy - A social psychological approach.* Boulder, CO: Westview Press.

De Houwer, J., Beckers, T., and Moors, A. (2007). Novel attitudes can be faked on the Implicit Association Test. *Journal of Experimental Social Psychology* 43**,** 972-978.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., and Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin* 135**,** 347-368.

De Waal, F.B.M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual review of psychology* 59**,** 279-300.

De Waal, F.B.M. (2012). "Empathy in primates and other mammals," in *Empathy: From bench to bedside,* ed. J. Decety. (Cambridge, MA: MIT press.), 87-106.

Eriksson, K., Andersson, P.A., and Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations* 19**,** 152-168.

Feldmanhall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., and Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition* 123**,** 434-441.

Frankenhuis, W.E., and Barrett, H.C. (2013). "Design for learning: The case of chasing," in *Social perception: Detection and interpretation of animacy, agency, and intentions,* eds. M.D. Rutherford & V.A. Kuhlmeier. (Cambridge, MA: MIT Press), 171-195.

Fujisawa, K.K., Kutsukake, N., and Hasegawa, T. (2006). Peacemaking and consolation in Japanese preschoolers witnessing peer aggression. *Journal of Comparative Psychology* 120**,** 48-57.

Gao, T., Mccarthy, G., and Scholl, B.J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological science* 21**,** 1845-1853.

Gao, T., Newman, G.E., and Scholl, B.J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology* 59**,** 154-179.

Gawronski, B., and Bodenhausen, G.V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin* 132**,** 692-731.

Gelman, R. (1990). First Principles Organize Attention to and Learning About Relevant Data: Number and the Animate-Inanimate Distinction as Examples. *Cognitive Science* 14**,** 79-106.

Gergely, G., and Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences* 7**,** 287-292.

Gonzalez-Liencres, C., Shamay-Tsoory, S.G., and Brune, M. (2013). Towards a neuroscience of empathy: Ontogeny, phylogeny, brain mechanisms, context and psychopathology. *Neuroscience and Biobehavioral Reviews* 37**,** 1537-1548.

Gredebäck, G., Kaduk, K., Bakker, M., Gottwald, J., Ekberg, T., Elsner, C., Reid, V., and Kenward, B. (2015). The neuropsychology of infants' pro-social preferences. *Developmental Cognitive Neuroscience* 12**,** 106-113.

Greenwald, A.G., Mcghee, D.E., and Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74**,** 1464-1480.

Greenwald, A.G., Nosek, B.A., and Banaji, M.R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology* 85**,** 197-216.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., and Banaji, M.R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97**,** 17-41.

Hall, L., Johansson, P., and Strandberg, T. (2012). Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE* 7**,** e45457.

Hamlin, J.K. (2013). Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core. *Current Directions in Psychological Science* 22**,** 186-193.

Hamlin, J.K. (2015). The case for social evaluation in preverbal infants: Gazing toward one's goal drives infants' preferences for Helpers over Hinderers in the hill paradigm. *Frontiers in Psychology* 5.

Hamlin, J.K., Wynn, K., and Bloom, P. (2007). Social evaluation by preverbal infants. *Nature* 450**,** 557-559.

Han, H.A., Czellar, S., Olson, M.A., and Fazio, R.H. (2010). Malleability of attitudes or malleability of the IAT? *Journal of Experimental Social Psychology* 46**,** 286-298.

Han, H.A., Olson, M.A., and Fazio, R.H. (2006). The influence of experimentally created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology* 42**,** 259-272.

Heberlein, A.S., and Adolphs, R. (2004). Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. *Proceedings of the National Academy of Sciences of the United States of America* 101**,** 7487-7491.

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

Hurlbert, A.C., and Ling, Y. (2007). Biological components of sex differences in color preference. *Current Biology* 17**,** R623-R625.

Johansson, P., Hall, L., Sikstrom, S., and Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science* 310**,** 116-119.

Kanakogi, Y., Inoue, Y., Matsuda, G., Butler, D., Hiraki, K., and Myowa-Yamakoshi, M. (2017). Preverbal infants affirm third-party interventions that protect victims from aggressors. *Nature Human Behaviour* 1**,** 0037.

Kanakogi, Y., Okumura, Y., Inoue, Y., Kitazaki, M., and Itakura, S. (2012). "Born to be sympathetic: Preverbal infants prefer victims and not aggressors", in: *XVIII Biennial International Conference on Infant Studies.* (Minneapolis, Minnesota).

Kanakogi, Y., Okumura, Y., Inoue, Y., Kitazaki, M., and Itakura, S. (2013). Rudimentary Sympathy in Preverbal Infants: Preference for Others in Distress. *PLoS ONE* 8**,** e65292.

Kenward, B., and Gredeback, G. (2013). Infants help a non-human agent. *Plos One* 8.

Koelkebeck, K., Hirao, K., Kawada, R., Miyata, J., Saze, T., Ubukata, S., Itakura, S., Kanakogi, Y., Ohrmann, P., Bauer, J., Pedersen, A., Sawamoto, N., Fukuyama, H., Takahashi, H., and Murai, T. (2011). Transcultural differences in brain activation patterns during theory of mind (ToM) task performance in Japanese and Caucasian participants. *Social Neuroscience* 6**,** 615-626.

Kuhlmeier, V.A. (2013). "The social perception of helping and hindering," in *Social perception: Detection and interpretation of animacy, agency, and intentions,* eds. M.D. Rutherford & V.A. Kuhlmeier. (Cambridge, MA: MIT Press), 283-303.

Kuhlmeier, V.A., Wynn, K., and Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science* 14**,** 402-408.

Larson, C., Aronoff, J., and Steuer, E. (2012). Simple geometric shapes are implicitly associated with affective value. *Motivation and Emotion* 36**,** 404-413.

Luo, Y., Kaufman, L., and Baillargeon, R. (2009). Young infants' reasoning about physical events involving inert and self-propelled objects. *Cognitive Psychology* 58**,** 441-486.

Lyons, A.B., and Cheries, E.W. (2017). Inferring Social Disposition by Sound and Surface Appearance in Infancy. *Journal of Cognition and Development* 18**,** 209-225.

Mar, R.A., Kelley, W.M., Heatherton, T.F., and Macrae, C.N. (2007). Detecting agency from the biological motion of veridical vs animated agents. *Social Cognitive and Affective Neuroscience* 2**,** 199-205.

Mathôt, S., Schreij, D., and Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44**,** 314-324.

Meissner, F., and Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology* 104**,** 45.

Meristo, M., and Surian, L. (2013). Do infants detect indirect reciprocity? *Cognition* 129**,** 102-113.

Mould, D., Mandryk, R.L., and Li, H. (2012). Emotional response and visual attention to non-photorealistic images. *Computers & Graphics* 36**,** 658-672.

Nosek, B.A., Smyth, F.L., Hansen, J.J., Devos, T., Lindner, N.M., Ranganath, K.A., Smith, C.T., Olson, K.R., Chugh, D., Greenwald, A.G., and Banaji, M.R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology* 18**,** 36-88.

Olson, M.A., and Fazio, R.H. (2004). Reducing the Influence of Extrapersonal Associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology* 86**,** 653-667.

Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P.E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *J Pers Soc Psychol* 105**,** 171-192.

Palmer, S.E., and Schloss, K.B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences* 107**,** 8877-8882.

Pratto, F., Sidanius, J., Stallworth, L.M., and Malle, B.F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology* 67**,** 741-763.

Premack, D. (1990). The infants' theory of self-propelled objects. *Cognition* 36**,** 1-16.

Premack, D., and Premack, A.J. (1997). Infants attribute value +/- to the goal-directed actions of self-propelled objects. *Journal of Cognitive Neuroscience* 9**,** 848-856.

Scholl, B.J., and Tremoulet, P.D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences* 4**,** 299-309.

Shamay-Tsoory, S.G., Aharon-Peretz, J., and Perry, D. (2009). Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain* 132**,** 617-627.

Simion, F., Bardi, L., Mascalzoni, E., and Regolin, L. (2013). "From motion cues to social perception: Innate predispositions," in *Social perception: Detection and interpretation of animacy, agency, and intentions,* eds. M.D. Rutherford & V.A. Kuhlmeier. (Cambridge, MA: MIT Press), 46-60.

Spelke, E.S., and Kinzler, K.D. (2007). Core knowledge. *Developmental Science* 10**,** 89-96.

Stevens, M., and Ruxton, G.D. (2012). Linking the evolution and form of warning coloration in nature. *Proceedings of the Royal Society B: Biological Sciences* 279**,** 417-426.

Takahashi, K., and Watanabe, K. (2015). Synchronous motion modulates animacy perception. *Journal of Vision* 15**,** 17-17.

Tatone, D., Geraci, A., and Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition* 137**,** 47-62.

Thomsen, L., Frankenhuis, W.E., Ingold-Smith, M., and Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science* 331**,** 477-480.