Eye-tracking studies on social attention have consistently shown that humans prefer to attend to other human beings. Much less is known about whether a similar preference is also evident in covert attentional processes. To enable a direct comparison, this study examined covert and overt attentional guidance within two different experimental setups using complex naturalistic scenes instead of isolated single features. In the first experiment, a modified version of the dot-probe paradigm served as a measure of covert reflexive attention toward briefly presented scenes containing a social feature in one half of the visual field compared to nonsocial elements in the other while controlling for low-level visual saliency. Participants showed a stable congruency effect with faster reaction times and fewer errors for probes presented on the social side of the scene. In a second experiment, we tracked eye movements for the same set of stimuli while manipulating the presentation time to allow for differentiating reflexive and more sustained aspects of overt attention. Supportive of the first results, analyses revealed a robust preference for social features concerning initial saccade direction as well as fixation allocation. Collectively, these experiments imply preferential processing of social features over visually salient aspects for automatic allocation of covert as well as overt attention.

Introduction

Social interactions are part of our everyday life and thereby essential for communication, empathy, and clarifying intent or needs for mutual understanding. Initial studies have confirmed a tendency to focus on other human beings in a social context, which has sparked extensive research in the field of social attention. Early studies frequently used isolated stimuli to compare allocation of attention toward social as opposed to nonsocial features (e.g., Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999; Pelphrey et al., 2002; Ricciardelli, Bricolo, Aglioti, & Chelazzi, 2002; Ro, Russell, & Lavie, 2001). One way to investigate this bias is by measuring the free allocation of eye movements to deduce attention distribution to these areas, also referred to as overt attention. However, some laboratory findings using simplified (e.g., iconic images or geometric shapes) or isolated stimuli (e.g., showing a face without the remaining body or contextual information) have yielded different results than studies using more naturalistic setups (Kingstone, 2009). When a face is presented in isolation, there is a clear preference to fixate the eye region of these images, but when a face is presented with its associated body parts, the tendency to scan the eyes disappears (Kingstone, Smilek, Ristic, etc.).
Friesen, & Eastwood, 2002). This altered behavior may be due to a lack of contextual information and situational complexity in the former as compared to the latter condition. Naturalistic scenes are rich in detail and provide meaningful contexts for objects influencing the selection process of eye movements, e.g., through scene gist (e.g., Oliva & Torralba, 2006) or a multitude of competing objects (e.g., Einhäuser, Spain, & Perona, 2008; but see Borji, Sihite, & Itti, 2013). This aspect is often diminished in impoverished visual displays (see also Anderson, Ort, Kruijne, Meeter, & Donk, 2015). As a consequence, researchers have started to utilize naturalistic scenes to extend previous results to a more realistically complex yet controlled setting. Recent studies have additionally introduced and investigated the presence of a human being within these scenes and its effects on gaze behavior (e.g., Birmingham, Bischof, & Kingstone, 2008a, 2008b; Cerf, Frady, & Koch, 2008; End & Gamer, 2017; Flechsenhar & Gamer, 2017; Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Rösler, End, & Gamer, 2017; Rubo & Gamer, 2018; Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014). These eye-tracking studies have shown that social features, such as faces or bodies of other human beings, are preferentially attended in free-viewing conditions. However, the use of naturalistic stimuli has mainly been restricted to experiments addressing overt attention by means of eye gaze but not for those that restrict eye movements in order to investigate covert shifts of attention. In addition to the distinction between covert and overt shifts of attention, the literature typically dissociates between reflexive, externally driven shifts of attention and voluntary shifts of attention. Traditional covert paradigms, such as variations of the Posner cueing paradigm (Posner, 1980) involve the presentation of cues that address either endogenous, top-down driven control or exogenous, bottom-up driven attentional orienting while restricting eye movements. In the original experiment, two boxes are presented on the screen, and one of these boxes will reveal a checkerboard. Before the onset of this probe, one of the boxes may be cued exogenously using a flash at the corresponding location, reflexively drawing attention to it. Alternatively, a central arrow cue may point to one of the boxes, triggering an endogenous shift of attention. The cues can either correctly indicate where a stimulus is going to appear (valid cue) or point to an incorrect location (invalid cue). Even when the cue does not reliably predict where the probe is going to appear, participants typically respond faster on trials in which the probe is congruent with the cue and, therefore, attended location compared to a cue that was incongruent. This phenomenon is referred to as a congruency effect. Although the majority of these paradigms use directional cues, such as arrows (e.g., Tipples, 2002), some studies have introduced social endogenous cues, referring to centrally presented faces or eyes indicating a certain spatial location by changing the gaze direction (Deaner & Platt, 2003; Driver et al., 1999; Friesen & Kingstone, 1998; Kuhn & Kingstone, 2009; Ricciardelli et al., 2002). Such gaze cues are able to induce automatic shifts of attention in the cued (i.e., gazed-at) direction and seem to resist voluntary control.

Dot-probe paradigms (MacLeod, Mathews, & Tata, 1986) make use of a similar principle to examine attentional biases. Participants fixate a central fixation cross and are presented with two peripheral stimuli for a brief amount of time. Subsequently, a probe appears at the location of one of these previously viewed stimuli, and participants are instructed to respond to it quickly, for instance, according to its location. If the probe succeeds the previously attended stimulus, reaction times will again be faster in line with the congruency effect. Herein, stimuli with emotional value (e.g., threatening words or angry faces) were shown to exert an attentional bias over simultaneously presented neutral stimuli. Although previous studies conducted in clinical setups (e.g., investigating patients with anxiety disorders) yielded largely consistent evidence in favor of an attentional bias for emotional faces (e.g., Bradley, Mogg, White, Groom, & Bono, 1999) as well as threat-related words (MacLeod et al., 1986; Mogg, Mathews, & Eysenck, 1992) or stimuli (Kroeze & van den Hout, 2000), findings in healthy participants have been less robust (see, e.g., Asmundson & Stein, 1994; Bar-Haim, Lamy, Bakermans-Kranenburg, & Van IJzendoorn, 2007; Mogg et al., 2000). Schmukle (2005) even concluded that dot-probe paradigms in general are inadequate to measure attentional biases in nonclinical samples as they have low test–retest reliability. However, Staagaard (2009) refined this statement by concluding that they might be unsuitable for individual-differences research but reliable for between-group designs to investigate different aspects of attention. Chapman, Devue, and Grimshaw (2017) addressed these inconsistencies directly, stating that electrophysiological evidence shows that healthy participants reliably attend to the emotional stimulus within such a dot-probe paradigm but that this may only be a brief occurrence, potentially causing reaction times to vary across trials leading to low reliability. Their results depicted greater reliability for stimulus onset asynchronies (SOAs, i.e., delay between the onset of the cue and onset of the probe) below 300 ms. Bindemann, Burton, Langton, Schweinberger, and Doherty (2007) also investigated different SOAs and simultaneously presented participants with a face cue and an object cue, one on each side of a central fixation cross. A subsequently presented probe that required a manual response (either detection or discrimination) was
presented at the same location as either the face or the object. When face and object cues were equally likely to predict a target, participants were faster to respond to probes that appeared in the same location as the face, suggesting that isolated faces capture covert attention. These results were largely independent of the SOA.

Although results of the previously mentioned studies investigating overt social attention using eye tracking (e.g., Birmingham et al., 2008a; End & Gamer, 2017; Fletcher-Watson et al., 2008) and covert social attention using spatial cueing paradigms (e.g., Bindemann et al., 2007; Kuhn, Tatler, & Cole, 2009) revealed comparable results regarding the preferential orienting toward social information, some researchers suggest that overt and covert shifts of attention are completely independent of one another (Posner, Cohen, & Rafal, 1982; Posner & Petersen, 1990) or are only partially overlapping (Corbetta et al., 1998). For instance, electrophysiological studies using steady-state visual evoked potentials (SSVEPs) have shown worse task performance (i.e., higher error rates and worse classification) and lower SSVEP amplitudes when eye movements were disallowed as opposed to when they were permitted (e.g., Kelly, Lalor, Finucane, & Reilly, 2004; Treder & Blankertz, 2010), suggesting that overt and covert orienting processes can be dissociated. Specifically, there seems to be a decrease in signal power in covert compared to overt attention (Ordikhani-Seyedlar, Sorensen, Kjaer, & Siebner, 2014). Overt attention reflected respective SSVEP frequency entrainment in the primary visual cortex, and covert attention initiated a shift toward parietal and frontal areas, that is, recruiting higher cognitive functions. Neuroimaging studies revealed inconsistent results as some attested to activation of identical brain regions during covert and overt orienting (De Haan, Morgan, & Rorden, 2008), and others emphasized a dissociation of involved brain regions: Cortical activations were similar for overt as well as covert attention, whereas the functional coupling changed as a function of task (goal-directed and stimulus-driven; Fairhall, Indovina, Driver, & Malalos, 2016). These deviating results raise questions about the dissociation between these two attention mechanisms. The previously mentioned studies that directly compared covert and overt attention mainly investigated general attentional mechanisms by using simplistic stimuli, such as geometric shapes (De Haan et al., 2008; Fairhall et al., 2016; Ordikhani-Seyedlar et al., 2014) or checkerboards (Kelly et al., 2004).

Only few studies have compared overt and covert attention in parallel in a social context (e.g., Domes et al., 2013) of which even fewer engaged naturalistic stimuli. An exception is the study of Kuhn et al. (2009), who manipulated overt attention using misdirection tricks in a video depicting a magician. However, the authors considered covert attention as a detection of the trick in this gaze-cueing paradigm, which only a few participants depicted (for similar approaches, see also Kuhn & Teszka, 2017; Kuhn, Teszka, Tenaw, & Kingstone, 2016). Thus, in order to further our understanding of the mechanisms of social attention, its overt and covert aspects should be compared more directly than in previous research.

The present study aims to provide such a comparison by addressing the following issues: First, we directly compared overt and covert social attention using identical stimuli in two independent studies in order to examine whether these aspects of attentional orienting are differentially sensitive to the presence of others (e.g., Kuhn et al., 2016; Risko, Richardson, & Kingstone, 2016). Second, our stimuli comprised naturalistic scenes containing a human being in different real-world situations in contrast to other studies (especially those investigating covert attention) that employed simplified (e.g., iconic images, geometric shapes) or isolated stimuli (lack of contextual information, e.g., face without a body). To the best of our knowledge, previous research has not yet examined covert shifts of attention toward social information within such more naturalistic viewing conditions. Finally, many studies indicating a social bias have not conclusively addressed the potential influence of low-level stimulus properties on attentional orienting (e.g., Castelhano, Wieth, & Henderson, 2007; DeAngelus & Pelz, 2009; Fletcher-Watson et al., 2008). Studies such as those of End and Gamer (2017) highlight the importance of accounting for low-level saliency, not only to rule out the possibility that a social bias is caused by higher visual saliency of the social features in a scene, but also to explain a worse prediction of gaze behavior by saliency in naturalistic scenes in which a social feature is depicted. Our stimulus set takes saliency distributions into account as we compared algorithms of center-surround, low-level feature predictions (Itti & Koch, 2000; Itti, Koch, & Niebur, 1998) as well as graph-based visual saliency (GBVS) predictions (Harel, Koch, & Perona, 2007), low-level isotropic contrast features (ICF; Kümmerer, Wallis, Gatys, & Bethge, 2017), and probabilistic models based on pretrained deep neural network features (DeepGazeII; Kümmerer, Theis, & Bethge, 2016) to account for low- and high-level image features in their efficiency to predict covert orienting as well as fixation patterns. In summary, we examined in two experiments whether social information is covertly and/or overtly selected over nonsocial information when embedded in real-world scenes.
Experiment 1: Covert social attention

In order to examine covert attentional orienting toward social information, we utilized the dot-probe task as an established experimental procedure. In the original version of the task (MacLeod et al., 1986), two cues are briefly presented simultaneously on the computer screen followed by a probe at one of the former cue locations that prompts a behavioral response. By manipulating cue categories (e.g., upright vs. inverted faces; see Langton, Law, Burton, & Schweinberger, 2008) and analyzing response times separately for cued and uncued locations, one can draw conclusions about which stimulus preferentially attracts attention. In the current study, we deviated from this standard procedure by presenting complex visual scenes that spanned the entire screen instead of separately stimulating both visual hemispheres. A single human being was present on one half of the picture (referred to hereafter as the social side of the stimulus), whereas the other half did not contain a human being (referred to hereafter as the nonsocial side of the stimulus). In the original version of the dot-probe task, a small dot is presented as the probe, and participants have to indicate its location by button press (MacLeod et al., 1986). This so-called detection variant has been criticized because participants could simply respond to the presence or absence of the probe at a specific location instead of distributing their attention to both locations (Bradley, Mogg, Falla, & Hamilton, 1998). Therefore, several studies have used a discrimination variant of the task in which participants do not simply respond to the probe location but instead have to respond to the identity of the probe, for example, by deciding whether the probe matches a pair of dots in the horizontal or vertical direction or whether a letter displayed as a probe is an “E” or an “F” (e.g., Chen, Ehlers, Clark, & Mansell, 2002). The current experiment consisted of both the detection and discrimination variant to determine attentional biases triggered by briefly shown (200 ms) naturalistic scenes. The detection task required participants to indicate the location of a presented probe (*), and in the discrimination variant, participants had to identify whether the letter “E” or “F” was presented at one of two locations. Participants were instructed to respond via keyboard to the probe as quickly and as accurately as possible while their eyes were to remain centrally fixated throughout the experiment. They were also aware of the fact that their response times were measured and their fixations controlled via a camera. By restricting eye movements through instruction to fixate a central cross, this setup allowed us to examine covert attentional orienting toward social information under naturalistic viewing conditions. Although the given tasks differed in complexity (discriminating letters being more demanding than detecting of a location), we hypothesized that an attentional bias toward the side of a depicted human being would result in a robust congruency effect for both manipulations, yielding faster reaction times and lower error rates for probes presented on the social side as opposed to the nonsocial side of the stimulus independent of task variant.

Methods

Participants

The study was conducted according to the principles expressed by the Declaration of Helsinki. The required sample size (n = 27) was estimated to allow for detecting a medium-sized advantage of the social stimulus side (d = 0.50) with a power of at least 0.80 at an alpha level of 0.05. To account for potential dropouts, we recruited 34 participants through the University of Würzburg’s Human Participant Pool and personal recruitment. Because all participants provided usable data, the final sample consisted of nine males and 25 females with a mean age of 21.85 years (SD = 2.72 years). Each participant had normal or corrected-to-normal vision, provided written informed consent, and received course credit for participation.

Apparatus

The experiment was programmed with Presentation® (Neurobehavioural Systems Inc., Version 18.1). Stimuli were presented on a 24-in. display (Fujitsu B24T-7 LED) with a size of 53.1 × 29.9 cm, resolution of 1,920 × 1,080 pixels, and refresh rate of 60 Hz. A chin rest with a forehead bar was mounted to the table and ensured a constant viewing distance of 72 cm from the eyes to the center of the monitor. A webcam (Microsoft LifeCam VX – 1000) was installed on the top of the display screen, and the resulting video image could be observed on a second computer that was not visible to the participant. The experiment was conducted in a secluded room with constant lighting conditions.

Stimuli

Stimuli consisted of 60 naturalistic social scenes, which were acquired through the Nencki Affective Picture System (Marchewka, Żurawski, Jednoróg, & Grabowska, 2014), EmoPics (Wessa et al., 2010), and Internet searches (e.g., Google picture search, Flickr). Scenes were cropped if necessary, contrast and luminance were adjusted manually to increase the homogeneity of the stimulus set, and image sizes were...
adjusted to 1,200 × 900 pixels (26.0° × 19.6° of visual angle). Each of the images contained a single person positioned on either the left or the right side of a complex naturalistic scene (for an example image as well as the overlap of social regions of interest for all images, see Figure 1). All original images could be flipped horizontally without compromising their appearance (i.e., stimuli did not include text or other elements that change meaning when mirrored). The distribution of visual saliency across image sides was controlled to avoid confounds of the location of social elements and low-level conspicuousness. Visual saliency maps were computed before the final selection of images using the procedure suggested by Itti and colleagues (Itti & Koch, 2000; Itti et al., 1998). This algorithm extracts three different low-level properties (i.e., orientation, intensity, and color contrast) to reveal locations that stand out from the background in terms of these features and combines them into a single topographic saliency map. Afterward, the final set of images was compiled such that 30 pictures had higher visual saliency on the social side (i.e., the half containing a person), and 30 pictures had higher visual saliency on the nonsocial side (i.e., the half opposite to the person) on the image. Six additional images were selected using the same criteria and served as a training set that was used to familiarize participants with the task.

**Paradigm**

In the current experiment, participants completed both a detection and a discrimination variant of the dot-probe task in separate blocks whose order was counterbalanced across participants. In each block, all 60 pictures were presented in their original orientation as well as in a horizontally flipped version, and each of these two versions was shown once with the probe on the social side and once with the probe on the nonsocial side of the stimulus. Thus, the side of the social feature, the relative visual saliency of the social side as well as the probe location were counterbalanced. This amounted to a total of 240 trials per condition. Trial order was randomized for each subject with the restriction that the same image could not appear in successive trials. After 120 trials, participants were given the opportunity to take a short, self-paced break.

In the detection variant, an asterisk (*) was used as probe, and participants had to indicate its location using their index fingers, which were placed on the keys “A” (left) or “#” (right) of a standard QWERTZ keyboard. In the discrimination variant, the letter “E” or “F” was used as probe. The probe was presented 6.6° of visual angle from the screen’s center. Both types of probes (asterisk/letter) were shown in white on a dark gray background in the font Arial with a size of 35 pt resulting in a visual angle of 0.4° × 0.4° for the asterisk and 0.6° × 0.8° for the letters. We randomly determined which letter was assigned to which trial but ensured an equal distribution of both letters within each set of original and flipped picture versions. Participants had to identify the letter using the keys “H” and the space bar of a standard QWERTZ keyboard using the thumb or index finger, respectively, of their dominant hand. The assignment of keys to letters was counterbalanced across participants for the discrimination task. The keys for both dot-probe variants were deliberately chosen in horizontal and vertical alignment to avoid confounds related to the Simon effect (Simon & Rudell, 1967), which results in shorter reaction times for

Figure 1. Illustration of stimulus characteristics. (A) Original scene with the human being on the left half of the picture (image taken from the Nencki Affective Picture System [NAPS], Marchewka et al., 2014, with permission for fair use). (B) Overlap of head and body regions of interest for all stimuli with the social information on the left half of the picture. Warm colors indicate high overlap of regions.
spatially congruent keys. Responses had to be given as quickly and accurately as possible.

**Procedure**

Participants were given task instructions and told to keep their eyes in the center of the display, indicated by a fixation cross at the beginning of each trial, at all times (covert viewing). It was pointed out to them that a webcam was installed on top of the monitor to ensure continuous fixation. The experimenter monitored the webcam video during the experiment in order to exclude participants with excessive eye movements. According to this visual assessment, all participants adhered to the instructions and showed no illicit eye movements during the task.

Each trial started with a fixation cross presented for 1 s, followed by an image that was shown for 200 ms. This presentation duration was chosen to avoid inhibition of return, which describes briefly enhanced attention at a cued peripheral location that is subsequently impaired for longer viewing durations (>255 ms; see Klein, 2000), but also to emphasize exogenous attention by reducing the influence of later (>300 ms, see Carrasco, 2011) top-down mechanisms. Directly after image offset, the probe appeared and remained on the screen until participants pressed a response key. After the behavioral response, the fixation cross reappeared in the middle of the screen for a random duration between 500 and 2,000 ms before the next trial started with the presentation of the fixation cross for 1 s and a subsequent stimulus presentation. The variable intertrial interval was chosen to avoid anticipation effects that may occur with consistent time intervals, enabling a certain preparation for the onset of the stimulus. For this reason, we also refrained from using a blank screen and upheld the presentation of a fixation cross to add to the unpredictability of the stimulus onset. Before starting with the detection or discrimination variant of the task, participants were familiarized with the procedure by completing 24 training trials with a different set of images. These trials were excluded from further analyses. Throughout the whole task, no error feedback was given, but the experimenter observed the participants’ responses during the training trials and gave advice if necessary (e.g., regarding the correct key assignment).

**Data analysis**

Reaction times and error rates were averaged per participant across trials separately for each cell of the 2 × 2 design consisting of the factors dot-probe variant (detection vs. discrimination task) and congruency (congruent vs. incongruent, i.e., probe presented on the social vs. the nonsocial side of the stimulus). Erroneous responses as well as outliers of response latencies were excluded from the calculation of mean reaction times. A value was defined as an outlier when it deviated more than 2.5 SD from the mean per subject and dot-probe variant. For the detection variant, this amounted to $M = 2.5\%$ ($SD = 1.0\%$) and for the discrimination variant to $M = 2.9\%$ ($SD = 0.8\%$) outliers. The influence of the experimental manipulations on error rates and reaction times was analyzed using 2 × 2 repeated-measures ANOVAs with the factors dot-probe variant, and congruency.

With our stimulus selection, we tried to control for systematic biases in visual saliency between the social and the nonsocial side of the images. Indeed, when relying on the algorithm of Itti and colleagues (Itti & Koch, 2000; Itti et al., 1998), visual saliency was balanced between both sides as a binary variable (i.e., the social side had a higher mean saliency in 50% of the images) as well as numerically with mean saliency values amounting to $M = 70.42$ ($SD = 15.81$) for the social and $M = 67.04$ ($SD = 20.99$) for the nonsocial side, $t(59) = 1.18$, $p = 0.24$, Cohen’s $d = 0.15$. However, recent algorithms proved to be more successful in the prediction of visual fixation patterns than the algorithm that was used here for stimulus selection (see, e.g., Borji & Itti, 2013; Judd, Durand, & Torralba, 2012). In order to ensure stability of our results for better performing algorithms, we did a post hoc analysis of our data using the GBVS algorithm (Harel et al., 2007), which was previously shown to outperform the Itti and Koch algorithm in the prediction of human gaze (Judd et al., 2012). Additionally, we incorporated the ICF model, which achieves top performance among all low-level baseline models (Kümmerer et al., 2017). Finally, we took into account the DeepGazeII model, which makes use of convolutional filters that have been pretrained on object recognition (Kümmerer et al., 2016). Herein, the algorithms based on low-level saliency (Itti, GBVS, and ICF) are considered cognitive models that use basic feature channels to create conspicuity maps, which are combined to generate predictions of gaze behavior in the form of saliency maps (Borji & Itti, 2013). DeepGazeII is considered a state-of-the-art model for saliency prediction using features from the VGG-19 deep neural network (Simonyan & Zisserman, 2014) pretrained on the SALICON data set (Jiang, Huang, Duan, & Zhao, 2015), showing high performance in the prediction of fixation patterns. The well-established MIT saliency benchmark offers a general ranking of saliency models (saliency.mit.edu; Bylinskii et al., 2016), and other comparisons were made with regard to information gain (quantification of the reduction in uncertainty of gaze allocation; see Kümmerer, Wallis, & Bethge, 2015; Kümmerer et al., 2017). Herein, the algorithms used in this study show increasing information gain in the following order: Itti,
GBVS, ICF, DeepGazeII (for an illustration of the visual saliency maps generated by these algorithms, see Figure 2B). Whereas the ICF algorithm did not reveal enhanced mean saliency on the social ($M = 25.01$, $SD = 8.80$) as compared to the nonsocial side of the stimulus ($M = 24.27$, $SD = 11.28$), $t(59) = 0.61$, $p = 0.54$, $d = 0.08$, different results were obtained for the GBVS as well as the DeepGazeII algorithm. These latter two models both revealed higher mean saliency on the social ($GBVS: M = 62.14$, $SD = 16.57$; DeepGazeII: $M = 24.74$, $SD = 11.28$) as compared to the nonsocial side ($GBVS: M = 56.77$, $SD = 15.90$; DeepGazeII: $M = 8.51$, $SD = 4.25$) of our stimuli: GBVS, $t(59) = 2.65$, $p = 0.01$, $d = 0.34$; DeepGazeII, $t(59) = 10.66$, $p < 0.001$, $d = 1.38$. In order to ensure that the current results were not merely driven by differences in the saliency distribution between the social and the nonsocial side of our stimuli, we repeated all analyses of the behavioral data for a subset of pictures for which the mean saliency was lower as well as the area of peak saliency (highest 5% of salience values) was smaller on the social as compared to the nonsocial side (GBVS: $n = 19$, ICF: $n = 24$, DeepGazeII: $n = 9$ images; see Figure 2D). This measure not only compares saliency of the social versus the nonsocial stimulus sides, but it also takes into account local saliency distributions that specifically drive early attentional selection.

All data processing and statistical analyses were performed using the statistical programming language R (version 3.2.3; R Core Team, 2016). An a priori significance level of $\alpha = 0.05$ was applied, and two-tailed testing was used throughout. For all ANOVAs, the generalized $\eta^2$ (Bakeman, 2005) is reported as effect size estimate.

**Results**

Reaction times for the full stimulus set (Table 1) were higher for the more challenging discrimination—as compared to the detection variant of the dot-probe task—resulting in a main effect of dot-probe variant, $F(1, 33) = 517.11$, $p < 0.001$, $\eta^2 = 0.692$. In addition, we observed a significant main effect of congruency, $F(1, 33) = 7.55$, $p = 0.01$, $\eta^2 = 0.001$, depicting longer reaction times for incongruent trials in which the social feature appeared on the opposite side of the probe. The interaction effect was not significant, $F(1, 33) = 0.03$, $p = 0.87$, $\eta^2 = 0.000$, implying that the facilitation effect of social content was similar in both dot-probe variants. These results were comparable when restricting the stimulus set to pictures that contained a nonsocial side with higher mean and peak saliency in terms of low-level features (Table 1). For the GBVS as well as the ICF algorithms, we obtained significant main effects of dot-probe variant, $F(1, 33) = 529.28$, $p < 0.001$, $\eta^2 = 0.692$ and $F(1, 33) = 493.41$, $p < 0.001$, $\eta^2 = 0.680$, respectively, and congruency, $F(1, 33) = 15.67$, $p < 0.001$, $\eta^2 = 0.009$ and $F(1, 33) = 16.45$, $p < 0.001$, $\eta^2 = 0.006$, respectively, but no statistically significant interaction of both factors, $F(1, 33) = 0.06$, $p = 0.81$, $\eta^2 = 0.000$ and $F(1, 33) = 2.07$, $p = 0.16$, $\eta^2 = 0.001$, respectively. However, for the reduced set determined by DeepGazeII, only a significant main effect of dot-probe variant emerged, $F(1, 33) = 464.35$, $p < 0.001$, $\eta^2 = 0.673$. Neither the main effect of congruency, $F(1, 33) = 2.38$, $p = 0.13$, $\eta^2 = 0.004$, nor the interaction of both factors, $F(1, 33) = 1.44$, $p = 0.24$, $\eta^2 = 0.002$, reached statistical significance. As depicted in Figure 3A, social information reliably modulated response times in both dot-probe variants, irrespective of its low-level visual saliency. When incorporating high-level feature predictions as modeled by the DeepGazeII algorithm, however, the congruency effect was no longer statistically significant. As indicated by the high variability of congruency effects in this condition, this might at least in part be related to the...
small number of suitable stimuli and, correspondingly, the small number of trials in the reduced stimulus set.

Similar to the reaction-time results, we also found a main effect of dot-probe variant on error rates for the full stimulus set, $F(1, 33) = 21.89, p < 0.001, \eta^2 = 0.082$, indicating a larger number of errors in the discrimination compared with the detection variant (Table 1). In line with the reaction-time data, we also obtained a main effect of congruency, $F(1, 33) = 10.57, p = 0.003, \eta^2 = 0.022$, depicting lower error rates for congruent

<table>
<thead>
<tr>
<th>DOT-PROBE VARIANT</th>
<th>RESPONSE TIMES (ms)</th>
<th>ERROR RATES (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Congruent</td>
<td>Incongruent</td>
</tr>
<tr>
<td>Full stimulus set</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection task</td>
<td>407.25 (53.15)</td>
<td>411.41 (51.72)</td>
</tr>
<tr>
<td>Discrimination task</td>
<td>592.69 (72.44)</td>
<td>597.39 (71.30)</td>
</tr>
<tr>
<td>Reduced stimulus sets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GBVS Detection task</td>
<td>404.97 (47.31)</td>
<td>415.95 (51.71)</td>
</tr>
<tr>
<td>Discrimination task</td>
<td>590.16 (68.94)</td>
<td>602.44 (76.10)</td>
</tr>
<tr>
<td>ICF Detection task</td>
<td>407.58 (52.36)</td>
<td>414.12 (53.02)</td>
</tr>
<tr>
<td>Discrimination task</td>
<td>590.03 (71.59)</td>
<td>603.91 (78.29)</td>
</tr>
<tr>
<td>DeepGazeII Detection task</td>
<td>405.28 (49.84)</td>
<td>419.76 (60.21)</td>
</tr>
<tr>
<td>Discrimination task</td>
<td>602.84 (79.28)</td>
<td>605.76 (77.30)</td>
</tr>
</tbody>
</table>

Table 1. Response times and error rates for the full as well as the reduced stimulus sets by algorithm (GBVS, ICF, DeepGazeII) as a function of dot-probe variant (detection vs. discrimination) and congruency of probe location and social information on the image (congruent vs. incongruent). Notes: Mean values across participants are depicted with standard deviations in parentheses. Reduced stimulus set: Values were calculated for a subset of 19 out of 60 stimuli for the GBVS algorithm, 24 out of 60 stimuli for the ICF algorithm, and nine out of 60 for the DeepGazeII algorithm that exhibited higher mean and peak visual saliency on the nonsocial side of the image.
trials as opposed to incongruent trials. Again, the interaction of both factors did not reach statistical significance, $F(1, 33) = 0.53, p = 0.47$, $\eta^2 = 0.001$. For the reduced stimulus sets as determined by the GVBS and ICF algorithms, only the main effect of dot-probe variant was found to be significant, $F(1, 33) = 7.29, p = 0.01$, $\eta^2 = 0.204$ and $F(1, 33) = 8.47, p = 0.006$, $\eta^2 = 0.204$, respectively. Neither the main effect of congruency, $F(1, 33) = 0.38, p = 0.54$, $\eta^2 = 0.001$ and $F(1, 33) = 0.85, p = 0.36$, $\eta^2 = 0.003$, respectively, nor the interaction of both factors, $F(1, 33) = 1.58, p = 0.22$, $\eta^2 = 0.007$ and $F(1, 33) = 0.45, p = 0.51$, $\eta^2 = 0.002$, respectively, reached statistical significance. For the reduced stimulus set as determined by the DeepGazeII algorithm, we did not obtain statistically significant effects: main effect dot-probe variant, $F(1, 33) = 1.32, p = 0.26$, $\eta^2 = 0.010$; main effect congruency, $F(1, 33) = 0.27, p = 0.60$, $\eta^2 = 0.001$; interaction of dot-probe variant and congruency, $F(1, 33) = 2.05, p = 0.16$, $\eta^2 = 0.013$.

**Discussion**

This study supports the notion that covert attention is preferably allocated to social features in complex naturalistic scenes. We observed reliably shorter reaction times and fewer errors for probes appearing on the side of the stimulus depicting a human being than for the nonsocial half, suggesting that social elements capture covert attention. Even though reaction-time differences between congruent and incongruent trials were relatively small for the whole stimulus set, this effect persisted for both variants of the social dot-probe paradigm with generally increasing demands of cognitive resources for the discrimination as compared to the detection variant. By carefully preselecting pictures with respect to the distribution of visual saliency and by restricting the data set to trials with higher mean and peak saliency on the nonsocial side of the pictures in post hoc analyses, we furthermore dissociated between influences of low- and high-level features in their efficiency to predict this social bias. Algorithms using low-level features only for saliency prediction demonstrated that preferential shifting of covert attention was largely independent of these image features. Moreover, the congruency effect in reaction times even seemed to be stronger for the reduced set of stimuli as compared to the full set, which discounts the influence of low-level saliency further. On these grounds, we suggest that social aspects are preferentially attended even when bottom-up influences should draw attention toward the opposite side. Algorithms involving a pretraining of deep neural networks seem to result in better predictions, confirming findings of Kümerer et al. (2017) that DeepGazeII performs particularly well on images containing faces. This further suggests that human gaze may be modeled more successfully using high-level features, especially with regard to reflexive social attention. However, it must be mentioned that the number of stimuli in the reduced set based on the DeepGazeII algorithm ($n = 9$) amounted to only 15% of the total data set, necessitating a reexamination of the effects elicited here. Nevertheless, our data also replicate and extend the findings of studies investigating preferential attention toward faces in comparison with other objects (e.g., Bindeemann et al., 2007; Devue, Belopolsky, & Theeuwes, 2012) and confirm that there is indeed a preferential selection of social information even when such features are irrelevant to the task at hand. Additionally, our findings illustrate that such effects can be observed even when using complex naturalistic stimuli instead of isolated or simplified ones (see Kingstone et al., 2002).

Although dot-probe paradigms have yielded consistent results regarding attentional biases toward emotionally salient stimuli in the past for clinical samples (Bradley et al., 1999; Kroeze & van den Hout, 2000; MacLeod et al., 1986; Mogg et al., 1992), findings in healthy participants have been shown to be less robust (see, e.g., Bar-Haim et al., 2007; Mogg et al., 2000). As mentioned previously, Schmukle (2005) even concluded that dot-probe paradigms in general are inadequate to measure attentional biases in nonclinical samples. Our paradigm differs from other dot-probe paradigms in several ways, somewhat refuting these doubts. First, based on the results of Chapman et al. (2017) that the reliability of attentional biases as measured by the dot-probe paradigm diminishes quickly with increasing SOA, we deliberately kept the SOA below 300 ms in the first experiment to specifically target initial and brief attention capture. Second, we did not investigate an attention bias toward threatening stimuli that may potentially generate an avoidance behavior in healthy subjects (Bar-Haim et al., 2007). Third, our paradigm did not use traditional directional or exogenous cues to guide attention that participants could have resisted by actively suppressing them when preparing for the upcoming probe presentation. In conclusion, the results of our experiment speak in favor of an early social bias for covert attention, which is evident without explicit cueing. Nevertheless, because our manipulation only allowed for automatic covert orienting, reflexive overt behavior as well as sustained attention processes should be further examined for this social stimulus set. In order to investigate this reliably and generalize the current findings, we conducted a second experiment using eye tracking with the same set of images to examine whether patterns of overt visual exploration are consistent with the observation of covert attentional orienting toward social cues.
In Experiment 2, a separate group of participants had their eye movements tracked as they freely viewed each of the scenes that were also used in Experiment 1. Participants were given no explicit task in this experiment and were instructed to freely view the stimuli like in a magazine. In order to differentiate between reflexive and more sustained aspects of visual attention, we showed half of the scenes for a very brief duration (200 ms) and quantified whether the first saccade was directed toward the social or nonsocial half of the image. The other stimuli were shown for longer durations (5,000 ms) to allow for a detailed exploration of the scene. Participants were aware of the fact that the images would be shown to them for different durations and were shown practice trials with a different set of stimuli before the experiment started. We hypothesized to observe a bias toward social information in reflexive attentional orienting as well as an increased exploration of faces in measures of sustained attention (see Birmingham et al., 2008a; End & Gamer, 2017; Rössler et al., 2017).

Methods

Participants

In order to account for potential dropouts due to low data quality, we recruited 33 participants (10 males and 23 females) with a mean age of 25.64 years (SD = 4.59 years). Participation was compensated in the form of course credit or monetary compensation. Because all data were of sufficient quality, the analyses are based on the whole sample of 33 participants. All other details were similar to Experiment 1.

Apparatus

Using Presentation® (Neurobehavioural Systems Inc., Version 18.1), stimuli were displayed on an LG 24MB 65PY-B (24-in.) screen (51.7 × 32.3 cm) with a resolution of 1,920 × 1,200 pixels and a refresh rate of 60 Hz. Eye movements were tracked in a dimly lit laboratory using a mounted EyeLink 1000 Plus system (SR Research Ltd., Ottawa, Canada). The camera had a 25-mm lens, and the system was set to a sampling rate of 1,000 Hz. Although viewing was binocular, only the right eye was tracked with a viewing distance of 50 cm from the monitor. Participants’ head position was stabilized using a chin rest and a forehead bar.

Stimuli

We used the same 60 stimuli (1,200 × 900 pixels) with counterbalanced low-level saliency on the social and nonsocial sides as in Experiment 1. With the current setup, the viewing angle amounted to 35.8° × 27.2° for each stimulus.

Paradigm

The experiment consisted of a 2 × 2 design with the within-subject factors presentation time (200 vs. 5,000 ms) and laterality of the social information (left vs. right). As described in the Methods section of Experiment 1, low-level visual saliency according to the algorithm of Itti and colleagues (Itti & Koch, 2000; Itti et al., 1998) was balanced across stimuli. Each image was only shown once during the experiment, and the association of each stimulus to the experimental conditions was determined randomly for each participant with an equal number of stimuli in each cell of the 2 × 2 design. Binary low-level saliency was balanced across presentation times such that half of the images had a higher saliency on the social and half on the nonsocial side. The order of all 60 trials was also randomized with the restriction that the same presentation time was not used in more than three successive trials to reduce expectation effects, which might potentially encourage rapid saccade initiation for short stimulus presentations or a lack thereof for consistently long presentation durations.

Procedure

Participants were told that they would be shown a number of images following a fixation cross. They were instructed to look at the fixation cross whenever present on the screen but that they were otherwise free to examine the images when presented. Before the experiment, participants were calibrated on the eye tracker using a nine-point calibration sequence. After successful validation, the trials began. Prior to each trial, a fixation cross was displayed for 1 s. The fixation cross was followed by the presentation of one of the stimulus images for a duration of either 200 or 5,000 ms. For brief stimulus durations, a uniformly gray screen was shown for 1,800 ms before the fixation cross reappeared again. Between trials, the fixation cross was presented for a randomly chosen period between 1,000 and 3,000 ms. Participants were familiarized with the procedure by completing six training trials (three with a 200 ms and three with a 5,000 ms presentation time) with a different set of images. These trials were excluded from further analyses.
Data analysis

Eye movements were parsed into saccades and fixations using thresholds of 30°/s for velocity and 8,000°/s² for acceleration for saccade detection. Fixations were defined as time intervals between saccades. A drift correction with reference to a baseline period of 300 ms before stimulus onset (i.e., during the presentation of the fixation cross) was accomplished for all fixations and saccades. To define outliers of baseline coordinates for fixation analyses, we used a recursive outlier removal procedure, which was applied separately for x- and y-coordinates (see also, e.g., End & Gamer, 2017; Flechsenhar & Gamer, 2017). Hereafter, the highest and lowest baseline coordinates were temporarily removed, and the mean and standard deviation were calculated for the remaining data for each participant. An interval bounded by 3 SD from the mean was set, and if any of the two temporarily excluded values fell outside of this interval, the value was removed permanently. If a data point fell within the interval, it was returned to the data set. This procedure was continued until no more data points were discarded permanently. Subsequently, the baseline position data of all scenes including a removed x or y baseline coordinate or missing baseline data (117 trials in total, average number of trials per participant: \( M = 3.54, \) SD = 3.69) were replaced by the individual mean of all scenes with a valid baseline position.

Three different analyses were carried out on these data. First, we analyzed the direction of the first saccade after stimulus onset as a measure of reflexive attentional orienting (see also, e.g., Rösler et al., 2017; Scheller, Büchel, & Gamer, 2012). In general, saccade latencies varied between 1 and 3,373 ms after stimulus onset (\( M = 328.61 \) ms, SD = 308.91 ms) with a minority of 26.48% starting before 200 ms. Thus, pictures in the condition with a short presentation time only served to trigger saccades that mostly occurred after stimulus offset when a uniformly gray screen was shown. In order to ensure that saccades were related to initial processing of the stimulus, we only scored saccades when their latencies ranged between 150 and 1,000 ms (cf. Rösler et al., 2017). Participants made saccades on the majority of trials (merely 4.72% of all trials did not contain saccade information) even though participants did not receive explicit instructions to initiate eye movements rapidly. Trials were excluded from the analyses when an invalid baseline was detected (i.e., when we could not ensure that participants fixated the center of the screen at stimulus onset) or when saccades were missing or occurring outside the scoring interval of 150 to 1,000 ms (238 invalid trials in total, average number per participant: \( M = 8.58, \) SD = 6.12).

Thereafter, the saccades were classified according to whether they were directed toward the social side of the stimulus (congruent saccades) or the nonsocial one (incongruent saccades). We then analyzed the proportion of congruent saccades as a function of presentation time (200 vs. 5,000 ms) using an ANOVA. For this analysis, we subtracted 50% from the respective proportions such that a significant intercept in the ANOVA model indicates a deviation from a chance distribution of saccade directions.

Second, we analyzed fixation densities as a measure of sustained visual attention for all trials with a presentation time of 5,000 ms (see also End & Gamer, 2017; Flechsenhar & Gamer, 2017). For these fixation analyses, trials containing too many blinks were excluded (192 trials in total with a blink-free time period of less than or equal to 80% of the whole trial, average number of excluded trials per participant: \( M = 5.82, \) SD = 5.26). Blinks in trials that contained more than 80% data were also removed so that the analysis of the remaining trials consisted of blink-free data. Fixation maps were generated for each participant and stimulus by adding fixations weighted by their fixation durations in milliseconds to an empty two-dimensional matrix (1,200 × 900 pixels). The first fixation was excluded when overlapping from the baseline period. The resulting map was then smoothed with a two-dimensional isotropic Gaussian kernel with a standard deviation of 32 pixels or 1° of visual angle using the R package spatstat (version 1.47.0, Baddeley, Rubak, & Turner, 2015). The resulting smoothing kernel comprised a width of 2° of visual angle (i.e., 1 SD in the positive as well as the negative direction) to resemble the functional field of the human fovea centralis. Finally, fixation density maps were normalized to a range from zero to one. In order to investigate the distribution of fixations onto the social and nonsocial features of the stimuli, we introduced four regions of interest (ROIs) similar to our previous studies (see also End & Gamer, 2017; Flechsenhar & Gamer, 2017). The head and body regions were manually drawn using the image manipulation software GIMP (version 2.8.14; GNU Image Manipulation Program, The GIMP Team). Additionally, we defined ROIs for areas with lower and higher visual saliency. This was achieved by partitioning saliency maps as calculated by the GBVS (Harel et al., 2007), the ICF (Kümmerer et al., 2016), and the DeepGazeII algorithm (Kümmerer et al., 2016). Saliency values smaller than or equal to the eighth decile of the saliency distribution were defined as areas of lower saliency and the remaining image regions as areas of higher visual saliency. This calculation was only accomplished for image regions that had not already been assigned to the head or body ROI. Although the eighth saliency decile represents an arbitrary choice, this cutoff proved useful in previous studies and allowed for the identification of image regions without social information but high low-level saliency (End & Gamer, 2017; Flechsenhar & Gamer,
2017). Figure 2B and C illustrate the saliency distribution and the resulting ROIs for one example picture. For each participant and scene, we determined the relative amount each ROI was fixated by calculating the sum of fixation density values for each ROI divided by the sum of fixation density values for the whole scene. This proportion score was then normalized by dividing it by the size of the ROI (in pixels) to control for increased fixations onto larger image areas (cf. Birmingham, Bischof, & Kingstone, 2009a). The mean for this relative area-normalized sum of fixation density was calculated for each ROI and participant, and a one-way ANOVA with the factor ROI (head, body, high saliency, low saliency) was carried out to examine the influence of image regions on fixation densities.

Third, we analyzed the location of the first five fixations post–stimulus onset for long stimulus presentations. The relative frequency for each ROI was determined individually by dividing the frequency of fixations directed toward an ROI by the frequency that any ROI was fixated. Similar to the fixation densities, these scores were normalized with regard to the area of the respective ROI (see Flechsenhar & Gamer, 2017). The resulting data were analyzed using a $4 \times 5$ repeated-measures ANOVA with the factors ROI (head, body, high saliency, low saliency) and fixation number.

In general, data analyses were conducted similarly as in Experiment 1. First analyses included the full set of all 60 stimuli that had similar visual saliency on the social and nonsocial sides as determined by the Itti and Koch algorithm (Itti & Koch, 2000; Itti et al., 1998). In this case, ROI definition was based on the GBVS algorithm to allow for a comparison with our previous studies (End & Gamer, 2017; Flechsenhar & Gamer, 2017; Rössler et al., 2017). A second set of analyses only included a reduced set of scenes with lower mean visual saliency and smaller areas of peak saliency (5% highest saliency values) on the social stimulus side as determined by three different algorithms: the GBVS (Harel et al., 2007), the ICF (Kümmerer et al., 2017), and the DeepGazeII algorithm (Kümmerer et al., 2016). For these analyses, we not only relied on the visual saliency algorithms for stimulus selection but also used the respective saliency maps for ROI definition. Therefore, these analyses allow for a comprehensive examination of an attentional preference toward social information in competition with nonsocial image regions with high visual saliency.

All data processing and statistical analyses were performed using the statistical programming language R (version 3.2.3; R Core Team, 2016). An a priori significance level of $\alpha = 0.05$ was applied, and the generalized $\eta^2$ (Bakeman, 2005) is reported as effect size estimate. Huynh–Feldt’s $\varepsilon$ is reported for all repeated-measures ANOVAs containing more than one degree of freedom in the numerator to account for potential violations of the sphericity assumption. Significant effects in the ANOVA were followed by post hoc pairwise comparisons with Bonferroni-corrected $p$ values.

Results

First saccade

Reflexive attention was analyzed as the relative frequency of first saccades directed congruently toward the social half of the stimulus (Figure 3B). Results of the full stimulus set revealed a significant intercept, $F(1, 32) = 235.25, p < 0.001, \eta^2 = 0.854$, indicating enhanced orienting toward the social as compared to the nonsocial stimulus side. The main effect of presentation time was not statistically significant, $F(1, 32) = 0.41, p = 0.53, \eta^2 = 0.003$. Concerning the reduced stimulus sets, results for a main effect of presentation time also failed to reach statistical significance for all analyses across algorithms: GBVS, $F(1, 32) = 0.54, p = 0.47, \eta^2 = 0.008$; ICF, $F(1, 32) = 0.93, p = 0.34, \eta^2 = 0.013$; DeepGazeII, $F(1, 32) = 0.93, p = 0.93, \eta^2 = 0.000$. Concerning a proportion of congruent saccades above chance level, we found significant intercepts for the reduced set of the GBVS, $F(1, 32) = 28.61, p < 0.001, \eta^2 = 0.321$, and the ICF algorithm, $F(1, 32) = 73.79, p < 0.001, \eta^2 = 0.559$, but not for the DeepGazeII algorithm, $F(1, 28) = 3.29, p = 0.080, \eta^2 = 0.321$. Thus, irrespective of whether stimuli were shown only briefly or for a longer duration, social content quickly attracted attention even when its low-level visual saliency was lower than for the nonsocial content. When saliency was defined through pretrained deep neural networks, however, the number of congruent saccades dropped to chance level.

Fixation density

In order to investigate aspects of sustained visual attention, fixation densities were assessed and analyzed when stimuli were shown for 5,000 ms (see Figure 4). Analyses revealed a main effect of ROI for the full stimulus set, $F(3, 96) = 165.99, \varepsilon = 0.35, p < 0.001, \eta^2 = 0.788$. The reduced sets defined by the GBVS, $F(3, 96) = 107.07, \varepsilon = 0.36, p < 0.001, \eta^2 = 0.702$; the ICF, $F(3, 96) = 136.58, \varepsilon = 0.37, p < 0.001, \eta^2 = 0.754$; and the DeepGazeII algorithm, $F(3, 96) = 76.70, \varepsilon = 0.36, p < 0.001, \eta^2 = 0.637$, also yielded significant main effects of ROI. Post hoc tests indicated heads to be looked at significantly more than all other image regions (all $p$s < 0.001). Although there was no significant difference between bodies and areas of higher saliency, both regions received significantly more attention than areas of lower saliency (both $p$s < 0.001). These effects were
comparable between the full stimulus set and the reduced ones.

First fixations

To follow up on the direction of the first saccade, the first five fixations were analyzed concerning the preference to select certain ROIs. For the full stimulus set, we found a significant main effect of ROI, $F(3, 96) = 157.73$, $\varepsilon = 0.35$, $p < 0.001$, $\eta^2 = 0.665$, and a main effect of fixation number, $F(4, 128) = 30.13$, $\varepsilon = 0.63$, $p < 0.001$, $\eta^2 = 0.083$. An interaction effect between ROI and fixation number was also significant, $F(12, 384) = 30.08$, $\varepsilon = 0.23$, $p < 0.001$, $\eta^2 = 0.262$. As shown in Figure 5, fixations were primarily directed toward the head of the human being followed by areas of high saliency. This effect was most pronounced in the second
Discussion

In Experiment 2, we investigated whether viewers overtly attend to social information earlier and longer than to nonsocial information. By using the same stimulus set as Experiment 1, this analysis allowed us to directly compare covert and overt attentional orienting toward social information. Our data reveal two main findings: First, participants made significantly more first saccades toward the side of the image that included a social stimulus irrespective of the presentation time. This initial attentional orienting even for short viewing durations (200 ms) that do not permit for an active exploration of the stimulus indicates that social attention is reflexive (Rösler et al., 2017). Second, these saccades also predominantly landed on social features, specifically on heads of the depicted human beings during long presentations (5,000 ms). Saccade target selection is, therefore, not only biased toward the social side in the beginning of scene presentation, but also specifically targets heads when allowed to further explore the scene. This attention bias was evident even for stimuli in which the social half of the stimulus was less salient in terms of low-level features than the nonsocial half (reduced stimulus sets for GBVS and ICF, respectively) and could be demonstrated even when comparing social elements (heads) against areas of high saliency. Furthermore, we took into account high-level information by using a stimulus selection based on the pretrained deep neural network of the DeepGazeII algorithm. In this case, social information no longer predicted the direction of initial saccades. Our results thereby not only suggest that social information is overtly selected over nonsocial information largely independent of low-level saliency, but also that this social bias might be more adequately captured by deep learning models. This is also reflected in the initial fixations, which revealed similar patterns for algorithms based on low-level features (GBVS, ICF) with a high priority for attending heads throughout all five fixations post–stimulus onset, yet different results for the DeepGazeII algorithm in which the first fixation was mostly allocated onto areas of high saliency. This again indicates that the DeepGazeII algorithm performed better in predicting potential fixation targets than the algorithms considering only low-level features but only for very early attentional mechanisms because subsequent fixations depict a reverted superior attention to heads, similar to those of the low-level feature algorithms. This suggests that pretrained features may be more successful in explaining reflexive social attention although its accountability deteriorates for later fixations. Concerning more sustained attention reflected by fixation densities over longer presentation durations, the reduced set of the DeepGazeII algorithm yielded similar results as those of the low-level saliency algorithms. A social bias persisted even when the mean and peak saliency was lower on the social side of the stimulus. Furthermore, the small sample of images in the DeepGazeII reduced set calls for further examination as the current results do not necessarily generalize to a larger set of heterogeneous stimuli. Nevertheless, the comparison between saliency algorithms based on low- and high-level features offers interesting insights for future approaches in modeling social attention. Another interesting aspect for future analyses is the additional consideration of local saliency distributions as well as the location of semantic information. Our method of reducing the stimulus set according to mean and peak saliency accounts for local and more distributed peaks of saliency but could be extended by implementing a pattern-based framework and object categories to specify how various factors, such as semantic information, contribute to saliency and gaze behavior (see Xu et al., 2014).

Several studies have shown that low-level features are capable of predicting fixation locations well above chance level in naturalistic scenes (e.g., Foulsham & Underwood, 2008) with greatest accuracy for initial eye movements after stimulus onset (e.g., Parkhurst, Law, & Niebur, 2002, but see Tatler, Baddeley, & Gilchrist, 2005). Rösler et al. (2017) used a generalized linear mixed model to reveal that saliency was a significant...
predictor of gaze behavior for initial saccades in social settings although social content surpassed saliency as a predictor in the model. On the one hand, our findings support the latter result as we controlled for low-level saliency distributions across our stimuli and further selected those for which mean and peak saliency were lower on the social half of the stimulus than on the nonsocial half. On the other hand, our data further suggest the feasibility of using probabilistic models, including transfer learning from other networks. The DeepGazeII algorithm as well as its predecessor were shown to be sensitive to faces (Kümmerer, Theis, & Bethge, 2014) and spatially specific face and person detection (Kümmerer et al., 2016) due to an increase in explained information through a combination of pretraining with features from a convolutional network. This seems to be the case in our data as well in accounting for reflexive (direction of the first saccade) but not for sustained social attention (prevailing high fixation densities on heads). Regarding results gained from purely low-level driven characteristics, our data using the GBVS and ICF algorithms are in line with the study of Birmingham, Bischof, and Kingston (2009b), who showed that the low-level saliency model did no better at predicting first initial viewing behavior toward the eyes than would be expected by chance. As previously shown by End and Gamer (2017), who compared the predictability of fixation behavior through saliency for social and nonsocial naturalistic scenes, low-level features seem to allow for better fixation location predictions when considering scenes without any social content but perform worse for images containing human beings. We conclude that low-level saliency does not account for the attention bias to select social features, and high-level features that might be implicitly tuned toward social information by means of pretraining perform better. Our study further suggests that social attention in naturalistic scenes is reflexive, which confirms the findings of Rösler et al. (2017) and nicely ties in with the results of Experiment 1 demonstrating rapid covert orienting toward social information.

**General discussion**

This study investigated mechanisms of social attention within overt and covert viewing conditions of complex naturalistic scenes in two different paradigms. Results of both experiments revealed an attention bias toward social information. A social dot-probe paradigm addressing covert attention revealed faster and more accurate responses to probes presented on the side of the social feature (i.e., human being) in the scene, whereas a subsequent eye-tracking study tested reflexive and sustained overt attention in accordance with these probe response data. Results derived from gaze behavior in the eye-tracking experiment revealed first saccade allocation toward the side of the social feature. First-saccade direction offered a direct measure for early visual attention upon the onset of a scene. Those first saccades in the short presentation condition indicate reflexive attentional mechanisms because a stimulus duration of only 200 ms precludes a detailed visual exploration of the presented stimulus (see Rösler et al., 2017). Thus, initial attention was preferentially directed toward social features covertly as well as overtly. In the covert dot-probe paradigm, responses for the detection task were faster than those for the more cognitively demanding discrimination task. For longer presentation durations in which participants could freely view the scene for 5,000 ms, fixations were also preferentially directed to social features, especially heads. Importantly, results of both experiments remained consistent when the data set was reduced to stimuli with higher mean and peak saliency on the nonsocial side with regard to low-level features. High-level features, on the other hand, accounted well for the saliency of social features with respect to early attention processes but less so for sustained social attention. Interestingly, the congruency effect tested within the dot-probe paradigm was higher for the reduced sets as compared to the full set for both tasks, and first saccades exhibited lower congruency effects for the reduced sets than for the full set. The fixation densities for the head region of the displayed human being were slightly higher in the reduced sets of GBVS and ICF than for the full set. However, reaction-time results and eye movements may not be directly comparable as eye movements represent a direct measure of visual processing (Kirchner & Thorpe, 2006; Parkhurst & Niebur, 2003), and response times require additional response initiation (Thorpe, Fize, & Marlot, 1996). Nevertheless, there is a possibility that covert attention is more prone to interference by social attention, at least for early attention processes, as fixation densities over longer viewing durations seem to depict a higher social bias for the reduced stimulus sets of the low-level feature algorithms compared to the full set. Despite this interesting dissociation, the combination of our social dot-probe with eye movement measures (as also suggested by Bantin, Stevens, Gerlach, & Hermann, 2016, for example) and coherent results of both experiments reinforce one another, and allow us to draw conclusions about both overt and covert attention processes. Our findings are also in line with previous studies showing an overt attention bias toward heads of human beings (e.g., Birmingham et al., 2008a, 2009a; End & Gamer, 2017; Fletcher-Watson et al., 2008) but also underline findings of experiments on covert attention toward faces (Bindemann et al., 2007).
Earlier studies addressing saccade programming provide further evidence for a tight temporal and spatial coupling between the preparation of saccades and prioritized visual processing and speak against the ability to attend one location while simultaneously preparing a saccade to another (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Kowler, Anderson, Dosher, & Blaser, 1995). According to these accounts, attention selection precedes saccade initiation and determines the end point of a saccade (see Kowler et al., 1995), independent of whether they were directed by peripheral (Schneider & Deubel, 1995) or central cues (Deubel & Schneider, 1996). Huestegge and Koch (2010) later suggested that attentional disengagement always precedes oculomotor disengagement, a process that applies to both covert and overt orienting, thus further supporting our results. Collectively, these results suggest reflexive covert as well as overt orienting toward social features in naturalistic scenes even when visually salient nonsocial information is present.

A question that is frequently raised in the context of attention toward other human beings is that social attention may be a form of top-down process engaged by ascribed meaning. Social information may be selected because it is most relevant to the viewer and may be considered a task goal in monitoring a given situation. However, considering the low interference of top-down instructions with social attention (Birmingham et al., 2008a) even when this social bias is detrimental to solving the task (Flechsenhar & Gamer, 2017) suggests that social stimuli may be effective on their own, independent of bottom-up or top-down mechanisms (see also Kuhn et al., 2009). Alternatively, Mack, Pappas, Silverman, and Gay (2002) conclude that attention is only captured after meaning has been assigned. This sequential processing would impose a very high perceptual load, which is refuted by our finding that social features are selected reflexively within 200 ms in the absence of explicit top-down influences. Even though others found covert and overt orienting to be independent, for reflexive (Hunt & Kingstone, 2003b) as well as voluntary shifts of attention (Hunt & Kingstone, 2003a), our results speak in favor of a linked relationship, at least in a social context. Although we did not look into neural mechanisms of overt and covert attention, the current findings might nicely correspond with the premotor theory of attention; that is, covert shifts of attention are seen as unexecuted overt shifts that rely on the same mechanisms (De Haan et al., 2008; Rizzolatti, Riggio, Dascola, & Umiltá, 1987).

With respect to reflexive versus sustained attention, our findings are coherent with the results of Cooper and Langton (2006) depicting an attention bias toward most relevant social features for short presentation durations (100 ms). However, they differ for longer presentation times: Although our eye-tracking study showed that attention toward social stimuli was dominant for both short (200 ms) and long (5,000 ms) presentation durations, the results of Cooper and Langton demonstrated a reverse effect for long durations (500 ms) explained by inhibition of return (Posner et al., 1982; Theeuwes & Van Der Stigchel, 2006). Other dot-probe paradigms using facial expressions as attention capture have shown similar inconsistent results and earned criticism (see Schmukle, 2005). Deviations of our results compared to those of Bindemann et al. (2007) and Cooper and Langton may originate from using complex scenes instead of isolated features and, therefore, complicate a direct comparison. Investigating naturally occurring phenomena before moving into laboratory settings as stated by the cognitive ethology approach, may allow for more concurrent findings, especially in social contexts (e.g., Kingstone, 2009). For instance, the Posner paradigm had important implications for research of spatial attention by differentiating between exogenous and endogenous cues for allocation of attention (Posner, 1980). By modifying this paradigm, Friesen and Kingstone (1998) demonstrated a discrepancy by revealing that, contrary to results using laboratory-generated cues, naturalistic directional cues with social value, such as eye gaze, which endogenously cued a location, triggered reflexive shifts of attention. This would indicate that social attention also relies on a different neuronal network than the ones suggested for bottom-up (ventral network) and top-down (dorsal network) mechanisms (Corbetta & Shulman, 2002). Objects are said to attract attention more efficiently when they are also relevant, which reinforces this assumption (Corbetta & Shulman, 2002). This demonstrates that an important challenge for future research is to reveal the neural underpinnings of social attention (cf. Adolphs & Spezio, 2006; Birmingham, Cerf, & Adolphs, 2011; Gamer & Büchel, 2009) and to define their organization with respect to other attention networks.

A number of mental disorders are particularly characterized by deficits in social functioning (e.g., social phobia or autism spectrum disorders). The dot-probe paradigm was initially conducted to examine vigilance and avoidance behavior in such clinical populations (e.g., MacLeod et al., 1986), but eye-tracking studies have also added to the existing literature. For example, patients with social phobia show an initial hypervigilance toward social features (Boll, Bartholomaeus, Peter, Lupke, & Gamer, 2016), sometimes followed by avoidance behavior (Horley, Williams, Gonsalvez, & Gordon, 2003), whereas patients with autism spectrum disorder show general avoidance behavior to social stimuli, especially to the eye regions of faces (Fletcher-Watson, Leekam, Benson, Frank, & Findlay, 2009; Klin, Jones, Schultz,
Volkmar, & Cohen, 2002). When given the choice between a face and an object, patients with social phobia were found to prioritize the object and thereby show the opposite effect of healthy cohorts who select social information eagerly (e.g., Chen et al., 2002). Similar patterns have been found concerning gaze following in patients with autism (e.g., Baron-Cohen, Campbell, Karmiloff-Smith, Grant, & Walker, 1995; Baron-Cohen, Wheelwright, & Jolliffe, 1997). Although our study demonstrates that covert and overt shifts of attention toward social information are highly connected in healthy individuals, it would be very interesting to examine their relationship and possible dissociation in patient groups.

**Conclusions**

The present study showed an attention bias for social features in complex naturalistic scenes that was evident for covert as well as overt attention. Many dot-probe paradigms have been previously utilized, but we are not aware of an existing social dot-probe paradigm like the one designed for this study. Even though we did not use cues, the social feature functioned as a cue by accelerating reaction times for probes presented on the social half of the stimulus. Additionally, we compared this dot-probe task to eye-tracking results in order to investigate covert and overt mechanisms of attention to the same set of stimuli and revealed a robust effect of preferential attention toward social features even in competition with highly salient nonsocial features. This social bias was evident for reflexive as well as sustained attention and highlights the importance of social features that are attended even in complex real-world scenes consisting of an abundance of competing, potentially more salient objects. These findings offer further support for the special role of social attention and set a strong example of social dominance.

*Keywords:* social attention, overt attention, covert attention, visual saliency, naturalistic scenes, eye movements, dot-probe task

**Acknowledgments**

The authors particularly thank Matthias Kümmerer for the provision and kind, helpful assistance regarding the implementation of his ICF and DeepGazeII algorithms. We further thank Eva Baur for her help in stimulus preparation and data collection. This work was supported by the European Research Council (ERC-2013-StG #336305).

Author contributions: M.G. and A.E. designed the study. O.L. collected data, A.F. and M.G. analyzed the data. M.G. supervised data analysis. A.F., O.L., A.E. and M.G. wrote and reviewed the manuscript.

Commercial relationships: none.

Corresponding author: Aleya Flechsenhar.
Email: aleya.flechsenhar@uni-wuerzburg.de.
Address: Department of Psychology, Julius Maximilians University of Würzburg, Würzburg, Germany.

**Footnote**

1 As participants saw each figure twice for every condition, we conducted an explorative post hoc analysis to investigate potential familiarity effects that may be reflected in reaction times. The $2 \times 2 \times 2$ repeated-measures ANOVA incorporating the factor repetition in addition to the factors dot-probe variant and congruency did indeed reveal a main effect of repetition, $F(1, 33) = 9.35, p = 0.004, \eta^2 = 0.002$, indicating that reaction times were faster when the image had already been presented before. Importantly, however, we obtained a main effect of congruency, $F(1, 33) = 8.33, p = 0.007, \eta^2 = 0.001$, but no statistically significant interaction between repetition and congruency, $F(1, 33) = 0.79, p = 0.378, \eta^2 < 0.001$. Thus, response times were reliably smaller on congruent trials, and this pattern was not significantly affected by familiarity.

**References**


