Social interaction detection by computer models in movement trajectories

Author: Lukas Graff
Supervisor: Dr. Ruud Hortensius
Honours supervisor: Prof. Daniel Cohnitz
Study programme: Artificial Intelligence (BSc), Utrecht University
Date of submission: July 2nd, 2021

– A 15 ETCS Bachelor thesis –

Content

1. Abstract

Even when watching animations of two dots moving on a white background, people often easily and automatically interpret what they see in terms of social interactions. Although the mechanisms by which people detect social interactions are not fully understood yet, recently, different cognitive computer models have been developed that fit human data well. The research question of this study is whether a bottom-up, hierarchical model designed by Shu et al. (2018) based on subinteraction classification and temporal interactivity parsing can explain human social interaction from movement trajectories. We designed a simple computer game which we used to generate stimuli trajectories and animations. In a second experiment participants rated the perceived social interaction in these animations. In line with our expectations, the hierarchical model fitted these experimental data well. However, average proximity between objects served as an almost equally good indicator for social interaction and an exponential model based on average proximity performed even slightly better. Using the principle of Occam's razor, we conclude that in videos of moving shapes, a simple proximity based approach offers a better explanation for human social interaction detection than the proposed hierarchical model.

Keywords: social interaction, animacy, social animations, cognitive modelling, social motion perception, interactive trajectory generation.

2. Introduction

Imagine walking through the dark when suddenly from the corner of your eye, you see something moving. It might be a mouse or a rat. All of your attention is drawn towards the object, until you notice it is just a leaf moved by the wind and you quickly lose interest. This example illustrates the apparent importance of detecting social objects or agents. Our brain and cognition appear to prioritize social scenes (Su et al., 2016) and to extensively process many different kinds of social interaction (Frith, 2007, 2008). In this study we aim to evaluate cognitive mechanisms that might be used in social interaction detection from movement trajectories by considering a computational model of this process.

Since a consensus on the definition of social interaction has not yet been reached, the following working definition of social interaction is used in this study: "We speak of a social interaction if two or more agents meaningfully and intentionally react to each other's behaviours." For our current study this definition suffices . As it turns out, in practice humans have a strong intuitive grasp of what social interaction entails. They can easily indicate whether a scene shows social interaction and they are strongly inclined to describe even the most primitive videos in terms of social interaction (e.g. fighting, chasing, dating). This was shown first by Heider and Simmel (Heider & Simmel, 1944). In their experiment participants were shown a very basic animation of moving geometrical shapes (see figure 1). The motion of these shapes was orchestrated in such a way as to suggest social and environmental interaction. These simple animations proved to be sufficient to evoke anthropomorphised descriptions of the shapes, often in terms of social interaction, even when participants were not explicitly asked to use such terms.

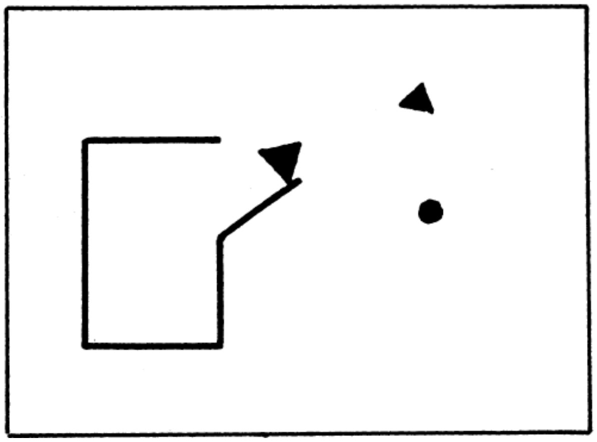


Figure 1: A frame of the original video as used by Heider and Simmel (1944).

There is a common set of biologically inspired subcategories of social interaction that many authors use, including: chasing (or pursuing), courting (or flirting), fighting, following, guarding and playing. (Barrett et al., 2005; Hovaidi-Ardestani et al., 2018; McAleer & Pollick, 2008). These categories can be explained quite well to people and can hence be used to generate stimuli depicting social interaction.

Another term often used for the social interaction perceived in stimuli similar to the Heider-Simmel animation is animacy. Although some studies on animacy focused on isolated agents / objects (Chandler-Mather et al., 2020; Tremoulet & Feldman, 2000), most studies focussed on animacy detection in stimuli containing at least two agents (Hovaidi-Ardestani et al., 2018; Isik et al., 2020; Schultz et al., 2005; van Buren & Scholl, 2017). In such a setup animacy and social interaction often go hand in hand and are hardly separable. If two geometrical shapes appear to move completely independently, they are not likely to be perceived as being engaged in social interaction, and neither as being animate, since animate beings do not neglect their surroundings. Similarly, if these shapes appear to react on each other's movements they are likely to be perceived as socially interacting as well as animate, since humans are only used to animate beings interacting socially. Indeed researchers found that cues clearly associated with social interaction (such as correlated movement and visibility of a second agent interacting with a first agent), increased animacy detection ratings as well (Dittrich & Lea, 1994; Schultz et al., 2005), while others stressed that animacy is a necessary condition of social interaction (Stewart, 1982). Hence we will build upon research about social interaction detection as well as animacy detection in animations showing movement trajectories of at least two objects / agents.

In several papers, researchers Gao and Scholl use experiments centred around animations showing the movement trajectories of at least two agents, to argue that perception of animacy is a low-level, automatic, non-introspectable mental process, "strongly and directly controlled by specific and subtle features of the visual input itself" (Scholl & Gao, 2013). In other words, they stress that animacy detection really is a perceptual process than a high-level cognitive process. In order to explore the presumed perceptual nature of animacy detection, Gao and Scholl prefer to focus on implicit behaviour measures of animacy detection rather than explicit verbal descriptions or ratings. Because of the stimuli they use and the manipulations they use to increase animacy perception, their concept of animacy coincides with the concept of social interaction.

In an experimental setup by Gao and Scholl (Gao et al., 2009; Gao & Scholl, 2011), called the *Don't-get-caught* task, participants controlled the movements of a "sheep" (represented by a dot), which moved on a screen with 20 other dots of a different colours moving seemingly randomly. One of these dots was a "wolf" and the participant's goal was not to be reached by this dot. The wolf dot did not stand out because of its colour. The only way in which it could be identified, was by its movements. If the wolf always moved directly towards the position of the sheep, participants performed well at escaping it. If on the other hand, the wolf was only generally directed to the sheep, but variated randomly within certain limits participants performed considerably worse (Gao et al., 2009). The same results were found when the wolf's chasing behaviour was interrupted by random movement (Gao & Scholl, 2011). Since sheep were able to move faster than wolves, a failure to escape from the wolf was explained as a failure to identify the wolf (rather than a failure in fleeing from it once identified). Hence performance in this experiment was interpreted as an implicit measurement of chasing behaviour (a form of social interaction). The strong dependence of this measure on subtle parametric differences was taken as evidence in support of the perceptual nature of animacy detection. Furthermore Gao and Scholl stressed that these findings could not be explained by reducing animacy perception to lower-level forms of perception, such as proximity perception (Gao et al., 2009; Gao & Scholl, 2011; Scholl & Gao, 2013).

Gao and Scholl also discovered and explored the wolfpack effect. This effect entails that when arrows (representing wolves) randomly move on a screen together with a dot representing a sheep, the orientation of the arrows influences animacy perception. When the arrows are always facing the sheep dot, they are more likely to be perceived as chasing the dot, as opposed to when they have a different direction (Gao et al., 2010). Gao and Scholl describe the wolfpack effect as a cue for animacy, but it could quite clearly (and perhaps even more convincingly) be described as a cue for social interaction. Again, Gao and Scholl support the perceptual nature of the detection of animacy cued by this effect by means of implicit measures. They also found lower performance in an adapted version of the *Don't-get-caught* task when distractors created the wolfpack effect as opposed to when they did not, which was taken as evidence that animate objects automatically draw attention. (Gao et al., 2010; Scholl & Gao, 2013). Using a matching pairs game paradigm it was also shown that stimuli in which the wolfpack effect is present are remembered better than stimuli without the effect (van Buren & Scholl, 2017).

All of these experiments seem to support the conclusion that animacy detection in animations showing the movement trajectories of at least two agents, is a form of perception rather than cognition. The perceptual nature of both animacy and social interaction detection is furthermore supported by research showing that participants of different cultures tend to categorize stimuli like those of Heider and Simmel in the same way (Barrett et al., 2005). A recent experiment explicitly tried to separate social interaction judgements from animacy perception. In this experiment, participants watched the original Heider and Simmel animation and had to press a button whenever they believed to perceive social interaction. From this experiment it was concluded that social interaction judgement is influenced both by low-level motion characteristics (such as distance between agents) and high-level understanding (Rasmussen & Jiang, 2019).

Neurocognitive studies however, seem to support the idea that social interaction processing might be a high-level, cognitive process rather than a low-level perceptual process. In one study, two Magnetoencephalography (MEG) signals correlating with social interaction detection and classification respectively were identified. Interestingly, these signals arose relatively late (300 ms for detection and 600 ms for classification) after stimulus onset. This was taken to suggest that social interaction detection and classification are high-level processes, involving top-down computation (Isik et al., 2020). Importantly however, the stimuli in this study were pictures of humans which were or were not interacting instead of videos of moving geometrical shapes.

Several neuroimaging studies have correlated the processing of animations of interacting shapes with activation in the posterior superior temporal sulcus (pSTS) and sometimes in the Temporo-Parietal Junction (TPJ) as well. (Isik et al., 2017; Schultz et al., 2005; Walbrin et al., 2018). The pSTS and TPJ are two neighbouring and not always clearly distinguished brain regions, which are hypothesized to play an important role in theory of mind processes, which involve reasoning about the perspectives and mental states of other agents (Aichhorn et al., 2006; Schurz et al., 2014, 2017). If this is true, it is unsurprising that these regions are activated when people watch simulations like those of Heider and Simmel, given that they tend to describe them in terms of social interaction and intentional actions. Although this evidence seems to point towards a high-level form of social cognition, this does not necessarily exclude the possibility of a more fundamental and primitive form of social interaction perception existing on a lower level. In fact, other research suggest that animacy cues in a most elementary form do not cause activity in the pSTS at all (Schultz & Bülthoff, 2019).

Another approach to researching human animacy and social interaction perception is found in the field of cognitive modelling. If a simple bottom-up model can be made that can account for human social interaction detection, this supports the hypothesis that human social interaction perception is a relatively low-level, bottom-up process. Two computer models for human social interaction detection from movement in animations like those of Heider and Simmel, featuring two objects or agents, have been developed, based on different underlying principles (Hovaidi-Ardestani et al., 2018; Shu et al., 2018).

The first model is a three-layer hierarchical model based on subinteraction classification and temporal interactivity parsing (Shu et al., 2018). This model is capable of making online, continuous predictions of the absence or presence of social interaction between two agents based on their movement trajectories. The first layer of this model consists of the motion pattern of a randomly chosen agent, relative to the other agent. In the second layer, the trajectory length is divided into time intervals of different lengths, each of which is assigned a latent subinteraction label (such as approaching, walking together or standing together), based on the motion patterns in the first layer. In the third and final layer of the model, each time interval is assigned a latent interaction label with a value of either 0 or 1, which represent the absence or presence of social interaction between both agents, respectively.

A Bayesian framework has been used to determine the joint probability of different subinteraction (second layer) and interaction label (third layer) sequences for a given motion trajectory (first layer). For each time point the model was able to predict the probability of presence of social interaction. This was done by multiplying the joint probability of each subinteraction label sequence and interaction label sequence by the value of the interaction label at that time point in that interaction label sequence (either 0 or 1). The outcome of this multiplication for each pair of subinteraction and interaction label sequences was added.

To train and test the model, aerial videos (captured by drones) of people in parks and on parking lots where used. From these videos motion, movement trajectories were obtained by manually marking the positions of two interacting persons in each frame. Of these trajectory pairs, 131 were used to train the model. The maximum number of interaction subcategories the model could identify was set to 15. All training stimuli were interactive, hence the model was trained without using negative examples. To validate and test the model, 24 interactive trajectory pairs were selected and 24 non-interactive stimuli were generated by interchanging trajectory-pairs. Furthermore, the drone videos corresponding to these trajectory pairs were used to create decontextualized animations only showing the position of two people as dots on a black background. These decontextualized videos could serve as stimuli in an experiment with human participants. In their experiment on human participants, Shu et al. (2018) used a setup in which participants could indicate in real time whether they perceived social interaction between two agents, did not perceive social interaction or were not certain. A high correlation ($r = 0.921$) and a low root-mean-square error (RMSE = 0.134) were found when comparing the model prediction to averaged human judgements. This suggest that this bottom-up hierarchical model which does not explicitly model intentions and goals is sufficient to explain human social interaction detection.

Another approach was taken by Hovaidi-Ardestani et al. (2018), who also created a hierarchical model to detect animacy and classify social interaction, but based their model on early neural visual processing. Hence, this model consists of a form pathway and a motion pathway, both of which in turn consist of a hierarchy of feature detectors. Each layer in this hierarchy is meant to (loosely) represent a layer of neurons in human visual processing, providing this model with a plausible biological grounding. Another, related reason why this model might be seen as more biologically plausible than the model by Shu et al. (2018)., is that it directly takes video data as input, instead of numeric motion trajectories. The influence of direction deviation, velocity change and the shape of the agents on the model prediction of animacy was qualitatively compared to the influence of these factors on human judgement and a similar trend was found. Furthermore, the model was able to correctly predict the type of social interaction in videos with an accuracy of 99.0%.

Research using behavioural measures suggests that social interaction detection in movement trajectories is a low-level process and shows some stimulus features that support this detection. It does not give clear mechanisms for this process. Neurocognitive research shows the activation of brain areas associated with high-level Theory of Mind processing, although this does not exclude the possibility of more fundamental processing taking place at lower levels. Finally, research using cognitive modelling approaches, supports the hypothesis of social interaction detection being a low-level, bottom-up process and suggest possible mechanisms for this process. For our research, we will turn our attention towards the hierarchical model designed by Shu et al. (2018) and ask the following question: can a hierarchical computational model explain human social interaction detection of moving objects? To test this, we will use a set of stimuli that are completely different from the ones used by Shu et al. (Shu et al., 2018), displaying a wide variety of interactive and non-interactive behaviour. Given the high similarity Shu et al. (2018) found between the prediction of this model and experimental participant ratings, our hypothesis is that this model can explain human social interaction perception. This would support the conclusion that social interaction detection is at least partially a low-level (perceptual) process.

Relevance to Artificial Intelligence

If the way humans detect social interaction from movement trajectories resembles models based on low-level visual processing, it is plausible that social interaction is a fundamental aspect of our visual perception. Much like depth perception, animacy perception is not something humans can turn off or have to construct actively. This would underline how sociality may be rooted deep inside human brains.

If the goal of AI is to imitate human intelligence, modelling human social interaction detection is an important step and can serve as a basis for emotional AI, a branch of AI which might have been slightly neglected, but which is increasing in popularity. On the other hand, the hypothesized low-level visual perception status of social interaction detection might be promising for the field of robotics. If social interaction detection is at base indeed a relatively simple process, even sensitive to moving shapes not intentionally controlled, it might be fairly easy to design robots which are perceived as socially interacting with each other or with humans. A realistic human or animal body and face might not even be needed as long as the robot's movements show social intentions.

Position within the field of Humanities

First of all Utrecht University has chosen to host their AI Bachelor program from their Humanities faculty. The reason behind this is that Utrecht University traditionally regarded AI as a philosophical discipline. All programming languages are grounded in formal logic (which originates in philosophy) and when studying AI, one inevitably encounters philosophical questions, such as "Are human behaviour and reason in theory ultimately reducible to processes of rule-following in the same way computer behaviour is?" Following this line of thought we might conclude that all AI research is positioned in the field of Humanities as well.

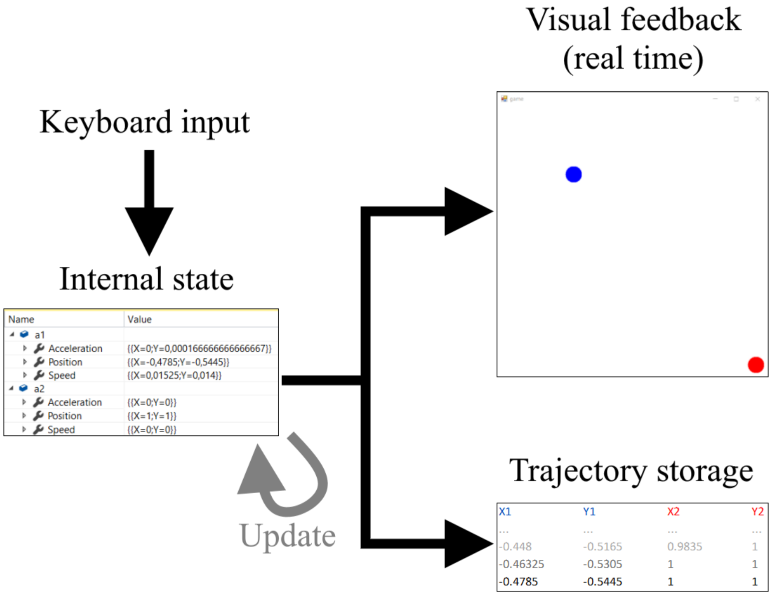
More specifically our research is about modelling human reason or cognition in a computer program. If our results prove this to be possible we showed that a particular aspect of human cognition is reducible to rule-following and hence contribute the hypothesis that all or most human cognition is reducible to rule-following. Furthermore our research can also have different impact on our view of humanity. If we find evidence that humans have no choice but to perceive particular movement trajectories as social interactions this supports the hypothesis that human beings are inherently social. The way we perceive and understand the world might be always influenced by our social cognition. This has great implications for our view of human kind, which positions this thesis not only within the field of cognitive and computer sciences, but also within the field of humanities.

3. Method

3.1. Trajectory generation experiment

We created a simple video game in which two participants used a computer keyboard to control the movements of a blue or red circle, generating movement trajectory pairs (lists of the positions of both agents). These trajectory pairs could in turn be used to (re)create animations, to extract features from and to serve as input for the hierarchical model by Shu et al. (2018).

Participants Six people (3 women, 3 men aged, 23 to 57 years old, average age 35, median age 25) participated in the trajectory generation task in pairs of two people. Each pair of participants was in a relationship and belonged to the same household. In this way, participants of the same pair did not have to keep 1.5 metres distance, which would be mandatory for participants from different households since the trajectory generation took place during the COVID-19 pandemic. Participants were acquaintances of the experimenter. Before the start of the experiment all participants received written information about the aims and content of the study and management of their data. They provided informed consent using an online form.



Name	Value
a1	
Acceleration	{{X=0;Y=0,000166666666666667}}
Position	{{X=-0,4785;Y=-0,5445}}
Speed	{{X=0,01525;Y=0,014}}
a2	
Acceleration	{{X=0;Y=0}}
Position	{{X=1;Y=1}}
Speed	{{X=0;Y=0}}

X1	Y1	X2	Y2
...	...	...	...
-0.448	-0.5165	0.9835	1
-0.46325	-0.5305	1	1
-0.4785	-0.5445	1	1

Figure 2: A schematic overview of the program used for stimulus generation. **Keyboard input:** Participants could use the keyboard to change the direction in which their agent accelerated and to increase their agent's maximum speed. **Internal state:** Here (among other things) the current maximum speed, position, the speed vector and acceleration vector of both agents were stored. Every moment before refreshing the frame, the internal state was updated. For both agents, the acceleration vector was added to the speed vector (as long as the length of the speed vector did not exceed the maximum speed value) and the speed vector was added to the position (as long as the agent would not collide with the other agent and the horizontal and vertical position coordinates of the agent would not be smaller than -1 or greater than 1). **Visual feedback:** Here the positions of both agents, as stored in the internal state, were visualized by dots on a white background. Additionally, larger goal circles could be shown. **Trajectory storage:** Every time the frame was refreshed, the positions of both agents were stored, resulting in a document containing the trajectories of both agents.

Design All pairs of participants were subjected to the same procedure and data collection was limited to a single session. Each pair of participants decided who would play as player 1 and who would play as player 2. This only effected the controls each participant had to use and the colour and starting position of the circle they controlled.

The social interaction game We used Microsoft Visual Studio 2017 to create the simple game participants played to generate trajectories and wrote additional code for manipulation and visualization of these trajectory pairs. The code is written in C#. Figure 2 summarizes the main idea behind the program. The code is available at https://gitlab.com/human-plus/sidcom/-/tree/main/Trajectory%20Generation%20Code

Participants were provided with images showing the real time positions of the objects each of them controlled. These objects were represented by a blue circle and a red circle on a white background. The size of each image was 600 by 600 pixels (see supplementary material A for some elaboration on the translation between coordinates and pixel indices). The diameter of both circles was 36 pixels. In some conditions, one or two goals were visible in each image, represented by circles with a diameter of 50 pixels. The colour of these circles was either a shade of blue or a shade of red, dependent on the agent the goal was intended for (see figure 3). The refreshment rate of these images was 60 frames per second (FPS).

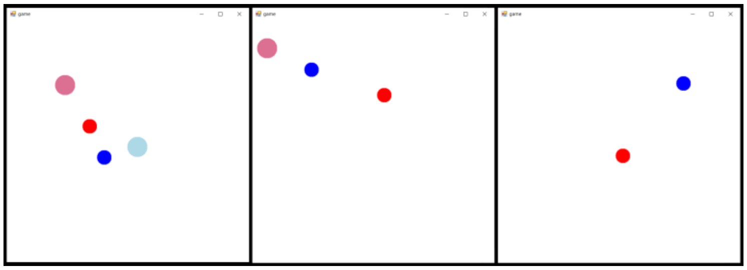


Figure 3: Three examples of possible frames during stimuli generation, containing either two goals, one goal or no goals.

Participants had to control the movements of the circle of their player colour by using a computer keyboard. Player 1 had to control a blue circle and player 2 had to control a red circle. See supplementary material B for the controls of both players. Each normal trial was preceded by a practice trial. These practice trials lasted for 25 seconds and could be repeated if participants felt they needed to. The trajectory pairs generated during practice trials were not stored. A normal trial lasted for 70 seconds and the trajectory-pairs generated during these trials were stored.

Procedure When participants started the program used for the generation of stimuli, a setup window popped up. Participants were instructed to choose a folder to which the trajectories they would generate would be saved, using this setup window, and they were asked to decide who would play as player 1 and who would play as player 2. Next, they were instructed how to choose different presets and how to select or deselect practice mode, using the setup window. Then, the controls were explained to both participants.

Next, participants repeated the following steps for six types of behaviour in a fixed order: 1) Selecting the preset corresponding to the relevant type of behaviour, 2) receiving instructions on the relevant type of behaviour, 3) practicing under the received instructions under the selected settings and 4) playing a “real” game under these instructions of which the generated trajectory-pairs were saved. We referred to those types of behaviour respectively as: free social interaction, chasing, individual goals, hindering, fighting and courting. Hindering and free social interaction are similar to the types of behaviour that have been referred to in other literature as guarding and playing, respectively. Before starting chasing, hindering and courting trials, participants had to decide which of them would be the chaser / hinderer / courter (both options required different presets). See supplementary material C for the instructions participants received before generating different social interaction categories.

3.2. Social interaction detection experiment

In this experiment we aimed to collect both explicit social interaction ratings and implicit social interaction perception measures from participants. Furthermore we obtained predictions based on models and stimulus features.

Participants In total, 34 people (23 women, 10 men, 18 to 71 years old, average age 25, median age 22) participated in this experiment. Of these participants, 8 were recruited via the Sona system of Utrecht University. (More information on https://techsupport.fss.uu.nl/labs/sona/). These were psychology students at the Utrecht University who received participation credits in return for their participation. The other 26 participants were acquaintances of the experimenter or supervisor and did not receive any reward in return for their participation. We excluded one participant from data analysis, because they provided a social interaction rating of zero for all stimuli, hence we used the data of 33 participants. Before the start of the experiment all participant received written information about the aims and content of the study and data management. They provided informed consent using an online form.

Design Each participant was randomly assigned to one of two variants of the experiment. The only difference between the variants was the set of stimuli participants were presented with. Of the participants, 15 participated in the first variant and 18 in the second. We chose to split the stimuli over two populations, because we wanted to limit the number of stimuli each participant was shown, while still being able to test a substantial number of stimuli. Limiting the number of stimuli for each participants was important for the feasibility of the recall task explained below.

Stimuli All stimuli were videos of a blue circle and a red circle moving on a white square. The diameter of the circles was 0.06 times the length of the square. On the left and right sides, the squares were surrounded by black borders of equal width, in order to obtain the common video ratio of 16:9. The webpages on which the stimuli were shown were white. The actual size of the stimuli was dependent on the device and browser settings of the participants. The length of all videos was 10 seconds and the refreshment rate of the videos was 60 FPS.

To obtain stimuli, we used the trajectory pairs that had been generated in the trajectory generation task. There were three trajectory pairs for each social interaction category (Free social interaction, Chasing, Individual Goals, Hindering, Fighting and Courting). Originally, each trajectory pair consisted of 4,200 entries, each entry containing the position of both agents at a specific point in time, coding a single frame. We deleted the first 600 entries of each trajectory pair, because these would be too similar, given the identical starting condition for each trajectory. We divided the remaining part of each trajectory pair into 6 sub-trajectory pairs of 600 entries each, coding precisely 10 seconds of video, given the framerate of 60 FPS. This resulted in 18 (sub-)trajectory pairs for each social interaction category. For each category, six trajectory pairs were randomly chosen to be used in the experiment. Furthermore 28 ‘shuffled’ trajectory pairs were created by combining two trajectories from different trajectory pairs not yet used in the experiment. We deleted shuffled trajectory pairs which would generate stimulus videos in which both circles would overlap and replaced them with new ones. We translated all 64 selected or created trajectory pairs into 10 second stimulus videos. Next, we assigned half of these videos to the first population of participants and the other half to them to the second population, preserving the distribution ratio over different interaction categories. For both populations, nine videos from shuffled trajectory pairs and two videos from trajectory pairs of the other interaction categories were selected to serve as stimulus videos in the rating task and targets in the recall task (see *task*). The other five videos of shuffled trajectory pairs and single videos of trajectory pairs of the other categories were selected to serve as distractors in the recall task.
Stimuli are available at https://gitlab.com/human-plus/sidcom/-/tree/main/stimuli

Task We used the website www.gorilla.sc to both design and conduct the experiment. Participants could access the experiment using their own personal computers or mobile devices.

Rating task: Participants watched a 10 second stimulus video, after which they were automatically shown a page on which they had to indicate the extent to which they perceived social interaction, using a slider widget. This screen also contained a progress bar showing the progress in the rating task. The slider gave an output on a scale from 0 to 100 which was divided by 100 to obtain a value between 0 and 1, which is the same scale used by the models we considered. Participants could not continue without rating the stimulus.

Recall task: Participants watched a 10 second stimulus video, after which they were automatically shown a page on which they had to indicate whether they recalled seeing this video during the rating task (yes/no). They could also press a button to go to this page before the video was finished. Each stimulus video was either a target, meaning that it was already presented during the explicit judgment task, or a distractor, meaning it was a novel video. Using signal detection approach, when participants indicated they recalled a target, this is called a hit; when they indicated they recalled a distractor, this is called a false alarm; when they indicated they did not recall a distractor, this is called a correct rejection; and when they indicated they did not recall a target, this is called a miss. Inspired by Van Buren & Scholl (2018) we aimed to use the miss/hit ratio of targets as implicit social interaction measure. Participants did not receive feedback.

Procedure Before starting the rating task, participants received an instruction, telling them that that they would be presented with animations featuring a red and a blue circle and that their goal was to determine whether these circles represented interacting social agents. It was stressed that 'social' did not necessarily mean co-operative in this context, but that we speak of social interaction if two or more agents meaningfully and intentionally react to each other's behaviours. After the instruction, the rating task started. Next, participants were informed about the recall task and participated in it. We did not explain the recall task earlier in order to avoid that participants would pay extra attention to all stimuli. After the recall task, participants had to indicate whether they had been distracted during the experiment. They were also asked to provide their gender and age, although they were allowed not to answer these questions if they wished to.

Additional data Besides the trajectories, stimuli and experimental data described above, which we will refer to as the game dataset, we have used another dataset. This dataset, which will be referred to as the drone dataset, consists of the trajectory pairs Shu et al. (2018) used in their experiment to test their hierarchical model and the experimental results obtained by them. Half of the trajectory pairs in this set originate from drone videos (only showing two positions as circles) and the other half were created by shuffling trajectory pairs (much like was done in this study). The experiment by Shu et al. (2018) had 33 participants, each of whom rated all 22 stimuli. Instead of providing one rating per stimulus (as was the case in our study), participants provided a continuous online rating of either 1, 0.5 or 0 by pressing different keys. We took the means of those ratings over time to obtain one rating value per video per participant.

Hierarchical model We also used the trajectory pairs which generated the stimuli videos for the experiment as input to the hierarchical model by Shu et al. (Shu et al., 2018). Because this model was tuned to another coordinate scale, we dilated and translated all trajectories. We also deleted three out of each four entries in each trajectory pair, because the model was tuned to a framerate of 15 FPS instead of 60 FPS. For each trajectory pair, the model's output was a list of predictions of social interaction ratings at different points of time during the trajectory. Each prediction had a value between 0 and 1. In order to compare these predictions with our experimental data, we took the average of these ratings per trajectory pair / stimulus video.

Proximity We also extracted some stimulus features from the trajectory pairs which generated the stimuli videos used in the experiment. The most important feature we extracted was the proximity between both agents, which is calculated by the formula:

$$\text{proximity} = \frac{\text{maximal distance} - \text{distance}}{\text{maximal distance}},$$

where distance is the distance between both agents. This proximity measure always has a value between 0 and 1, which is larger if the agents are closer together. In our analysis we used the mean proximity for each stimulus trajectory and in the following sections we will use the term proximity to refer to this mean.

Exponential proximity model During our initial data analysis, we discovered that there appeared to be an exponential rather than a linear relation between proximity and social interaction rating. Furthermore, social interaction predictions based on proximity generally seemed to be slightly lower than the actual ratings. Hence, we constructed the following simple model structure: $\hat{y} = \text{proximity}^{a} + b$, where $\hat{y}$ is the predicted social interaction rating and $a > 1$ and $b \geq 0$, are parameters that can be used to fit the model to specific data. Parameter a incorporates the notion of an exponential relation and b can be used to slightly increase all predictions.

We fitted the exponential proximity model on the drone dataset. First, we tuned parameter a by gradually increasing it with steps of 0.1 (starting at 1.1), until we found a value that maximized the Pearson corelation coefficient between $\hat{y}$ and the average experimental social interaction rating for each stimulus. Once we fixed a, we tuned b by gradually increasing it with steps of 0.01 (starting at 0), until we minimalized the root mean square error (RMSE) between $\hat{y}$ and the average experimental data.

Data Analyses In order to show whether the hierarchical model by Shu et al. (2018) is a plausible model for human social interaction detection, we compared the model's predictions with our experimental results. We did this by calculating the Pearson correlation coefficient (r), the squared the Pearson correlation coefficient (r^2) and root mean square error (RMSE) between the model prediction and the mean experimental social interaction rating for each stimulus. We used a t-test to determine whether the found correlation coefficients we found were significantly larger than 0. We also determined the fit of this model on the social interaction ratings of each individual participant by calculating the (squared) Pearson correlation coefficients and RMSE's between the model prediction and these individual social interaction ratings. These values were averaged to obtain a mean correlation coefficient ($\bar{r}$), mean squared correlation efficient ($\overline{r^2}$) and mean RMSE ($\overline{\text{RMSE}}$) for the model's performance on individual participant data, along with standard deviations (SD) of all of these values. Considering correlations and RMSE's for individual participant data has two main advantages: firstly, it shows whether a model can explain social interaction detection in individuals rather than averaged over groups and secondly it allows for a straightforward statistical comparison between different models, more on which below and in supplementary material E.

We did not just want to determine whether the hierarchical model performed well. To determine whether this specific model explains human social interaction detection, we needed to determine whether its prediction accuracy could be matched by a prediction based solely on simple stimulus features. From initial analysis we concluded that proximity was the most promising feature for predicting social interaction detection. Hence, we repeated the analyses described above for proximity and the fitted exponential proximity model described above versus experimental rating. Finally, we calculated the differences between the $\bar{r}$- and $\overline{\mathrm{RMSE}}$-values of all social interaction rating predictors we used (the hierarchical model, proximity and the exponential proximity model). We performed paired two-sample t-tests to determine the significance of these differences.

Additionally, we calculated the Pearson corelation coefficients between explicit social interaction ratings and hit/miss rates in the recall task, as well as the corelation coefficients between the model predictions and these hit/miss rates and proximity and these hit/miss rates. We performed t-tests to determine the significance of these correlation coefficients.

4. Results

Testing the hierarchical model on the game dataset The hierarchical model by Shu et al. (2018) fitted the experimental data from the game dataset well. In figure 4, we plotted the predictions of this model for each stimulus video versus the mean social interaction rating over all participants for the same stimulus. A qualitative analysis of figure 4 shows that stimuli in the shuffled and individual goals categories received the lowest social interaction ratings, both from participants and the hierarchical model, whereas stimuli in the fighting and courting categories received high ratings. We found a strong correlation with a relatively low RMSE between the hierarchical model and the experimental data (see table 1).

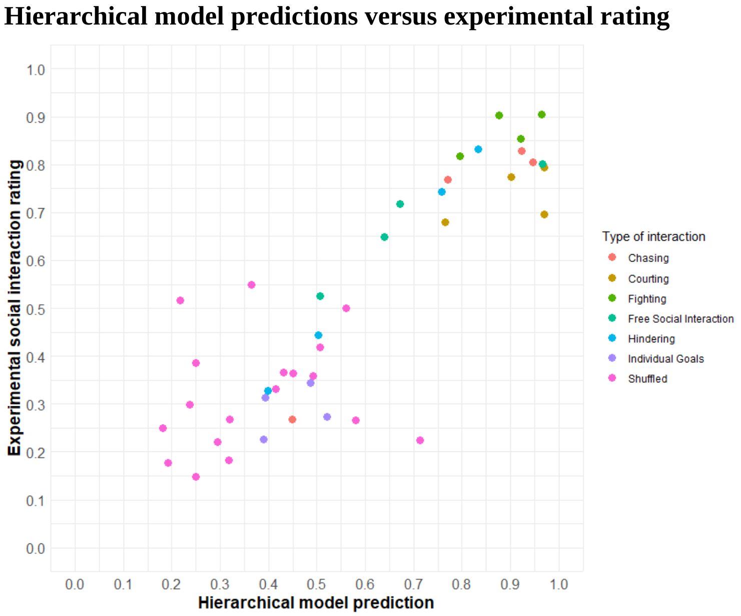


Figure 4: The predictions of the hierarchical model by Shu et al. (2018) versus the mean of the explicit experimental social interaction ratings of all participants in the game dataset. Each dot represents an animation. The colour of each dot shows the interaction label of this animation.

Fitting the exponential proximity model on the drone dataset To fit the parameters of the exponential proximity model, we used the drone dataset. We found an optimal fit for the formula $\hat{y} = \text{proximity}^{7.0} + 0.08$.

Performance of proximity as predictor and the exponential proximity model on the game dataset Figure 5A plots proximity against mean social interaction rating. Again, a qualitative analysis shows that the fighting and courting stimuli have the highest proximity and the shuffled and individual goals categories have the lowest proximity. We found that the exponential function $y = \text{proximity}^{5.6} + 0.16$, fitted the experimental data best. The curve resulting from this formula is plotted in figure 5A. This formula can also be interpreted as an exponential proximity model fitted on the game dataset instead of the drone dataset. Figure 5B plots the predictions of the original exponential proximity model, fitted on the drone dataset, versus the mean experimental ratings.

Summary of the performance of all predictors Table 1 provides a quantitative analysis of the performance of all social interaction predictors on the averaged social interaction ratings for both the drone and game dataset. All corelations are significantly higher than zero ($p < .01$). The table also contains the t-values of the correlations. Table 2A provides the average fit for all predictors on the social interaction ratings of each individual participant in the drone dataset. Table 2B provides the same data for the game dataset.

Predictor	**Drone dataset**	r	$t_r(20)$	r^2	RMSE	**Game dataset**	r	$t_r(40)$	r^2	RMSE
Hierarchical model		.96	15.60	.92	.10		.86	10.71	.74	.15
Exponential proximity model		.87	7.85	.76	.14		.93	15.98	.86	.18
Proximity		.84	6.98	.71	.34		.88	11.50	.77	.30

Table 1: A summary of the performance of different predictors for the social interaction rating means in both datasets. The drone dataset contained 22 stimuli and the game dataset contained 42 stimuli.

.

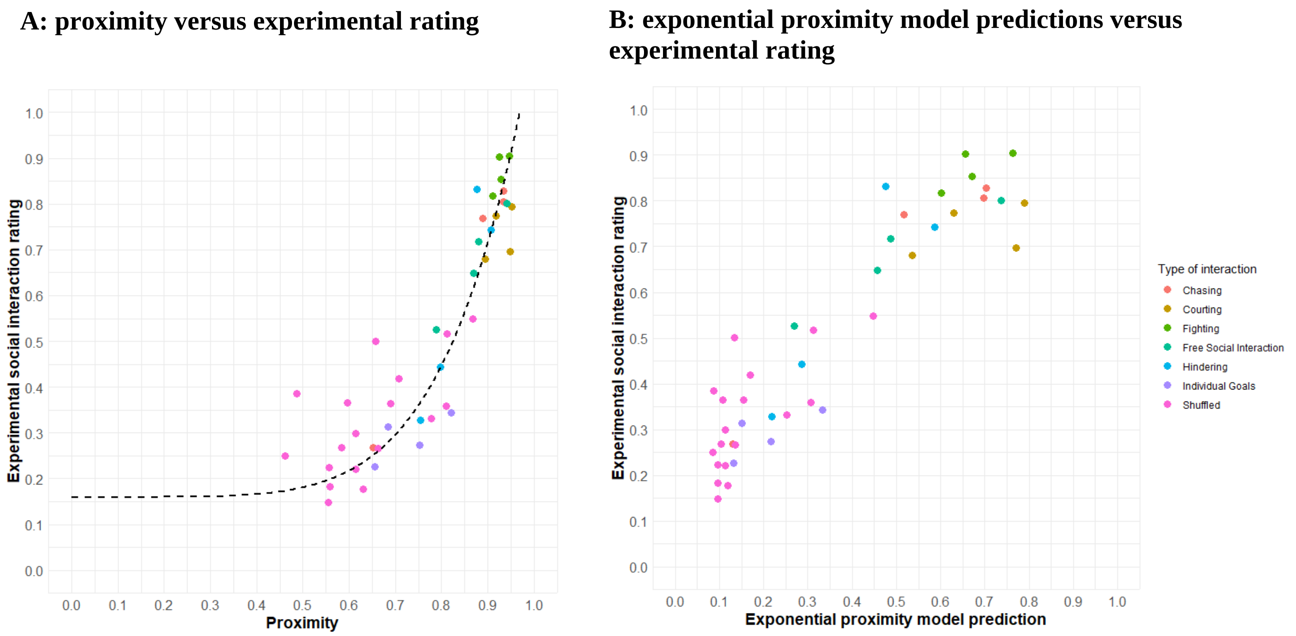
A: Averaged performance on individual data in drone dataset

Predictor	$\bar{r}$	SD_r	$\overline{r^2}$	SD_{r^2}	$\overline{\mathrm{RMSE}}$	$\mathrm{SD}_{\mathrm{RMSE}}$
Hierarchical model	.85	.19	.76	.15	.16	.052
Exponential proximity model	.77	.20	.63	.17	.19	.047
Proximity	.75	.18	.59	.16	.35	.044

B: Averaged performance on individual data in game dataset

Predictor	$\bar{r}$	SD_r	$\overline{r^2}$	SD_{r^2}	$\overline{\mathrm{RMSE}}$	$\mathrm{SD}_{\mathrm{RMSE}}$
Hierarchical model	.68	.17	.48	.17	.26	.088
Exponential proximity model	.74	.12	.56	.17	.28	.073
Proximity	.70	.18	.50	.21	.36	.13

Table 2A: The average prediction performance measures for all individual participant ratings in the drone dataset and their standard deviations (SD). **B:** The average prediction performance measures for all individual participant ratings on the game dataset and their standard deviations (SD).



A: proximity versus experimental rating

B: exponential proximity model predictions versus experimental rating

Figure 5: The mean proximity between both agents during a stimulus video versus the mean of the social interaction ratings of all participants for the game dataset. Each dot represents an animation. The colour of each dot shows the interaction label of this animation. The dashed black line represents the formula $f(x) = x^{5.6} + 0.16$, which fitted the data best. **B:** The predictions of the model fitted on the drone dataset, represented by the formula $\hat{y} = \text{proximity}^{7.0} + 0.08$, versus the mean of the social interaction ratings of all participants for the game dataset. Each dot represents an animation. The colour of each dot shows the interaction label of this animation.

Comparison between models Finally, we used the different values of $\bar{r}$ given in table 2 to compare the performance of the hierarchical model, the exponential proximity model and the proximity for both data sets. The results of this analysis are given in table 3. Using $\overline{\text{RMSE}}$ instead of $\bar{r}$ as measure of fit resulted in a similar outcome, except for the fact that the $\overline{\text{RMSE}}$ of the hierarchical model for the game dataset was significantly lower than the $\overline{\text{RMSE}}$ of proximity for the same dataset and was even lower than the $\overline{\text{RMSE}}$ of the exponential proximity model for this dataset, although this difference was not significant. (Note that a lower RMSE indicates a better fit.) See supplementary material D for a version of table 3 using $\overline{\text{RMSE}}$ instead of $\bar{r}$.

The difference in $\bar{r}$ for different predictors on both datasets.

	Dataset	Proximity	Exponential proximity model	Hierarchical model
Hierarchical model	Game dataset	− .017	− .061	
	Drone dataset	+ .10	+ .080	
Exponential proximity model	Game dataset	+ .044		+ .061
	Drone dataset	+. 023		− .080

Table 3: The differences between the mean Pearson correlation coefficients for all predictors. The numeric values in each cell represent the difference between the $\bar{r}$ resulting from using the row predictor and the $\bar{r}$ resulting from using the column predictor. The second column specifies the dataset used to obtain both $\bar{r}$'s. All original $\bar{r}$-values can be found in tables 2. Green cells indicate a significant positive difference ($p < .01$), orange cells indicate no significant difference and red cells indicate a significant negative difference ($p < .01$). To determine significance, paired two-sampled t-tests were used. Supplementary material E contain the corresponding t-values.

Recall experiment We found a weak, but just significant corelation between correct hit/miss ratio over all participants and proximity ($r = .32$; $t_r(40) = 2.10; p < .05$). We found slightly less weak, significant corelations between mean experimental social interaction rating and hit/miss ratio ($r = .45$; $t_r(40) = 3.23; p < .01$), hierarchical model predictions and hit/miss ratio ($r = 0.44$; $t_r(40) = 3.13; p < .01$) and exponential proximity model predictions and hit/miss ratio ($r = .45$; $t_r(40) = 3.19; p < .01$).

Distracted and interrupted participants Two participants in our experiment indicated the extent to which they were distracted during participation in the social interaction detection experiment to be higher than 50 on a scale from 0 to 100. Two other participants indicated they had been interrupted for longer than 10 minutes while participating. When excluding all four of these participants from our analyses we obtained similar performance measures for both models and main findings from table 3 (that is to say whether we found a significant difference in $\bar{r}$ and $\overline{\text{RMSE}}$ given two predictors and a dataset and whether this difference was positive or negative) did not change. See supplementary material F for a version of figure 4 generated with exclusion of these four participants.

5. Discussion

Main findings The present study has examined human social interaction perception and re-evaluated the claim that a hierarchical model based on subinteraction classification and temporal interactivity parsing can explain human social interaction detection in animations similar to the ones first used by Heider and Simmel (1944). We tested this model on a novel dataset and translated its continuous online prediction output to single values per participant per stimulus. We were able to confirm the claim by Shu et al. (2018) that their model achieves a high performance, even under these new conditions. We did find a considerable performance drop compared to the performance on the original testing data, however. We also found a moderate average corelation between the hierarchical model predictions and individual social interaction ratings.

However, solely on the basis of the high performance that was found, we cannot yet conclude that this hierarchical model provides a good explanation of human social interaction detection. There could be several social interaction detection models that match human performance equally well. For this reason, we also considered an exponential, proximity-based model. By testing on the game dataset no strong evidence that the hierarchical model is a better predictor for social interaction rating than proximity was found. The exponential proximity model even outperformed the hierarchical model in terms of average correlation and did not significantly underperform in terms of RMSE.

Moreover, the hierarchical model is rather complex, since it adds to the process of social interaction detection from movement trajectories the concept of subinteractions and many hidden variables. The exponential proximity model however, is simple, since its predictions are based on one straightforward calculation only using two parameters. According to the principle of Occam's razor, one should prefer the simplest, correct explanation, where simplicity in case of cognitive models is, among other things, determined by the number of parameters used (Myung & Pitt, 1997). Hence the simple exponential proximity model should clearly be preferred over the more complex and not clearly better performing hierarchical model and the hierarchical model does not provide the best explanation for human social interaction detection.

Furthermore, in our study we have introduced and designed a tool which can be used to create large amounts of Heider and Simmel stimuli relatively easily. We were not the first to take this approach (Barrett et al., 2005; Blythe, 1999), but we did not found another study which used a tool which did not only display the agents the participants could control, but also goals which agents had to move to. This, in combination with easily adaptable speed settings and different pre-sets, makes our tool suitable for creating a wide range of simple motion animations for many different purposes. The dataset that was generated by using this tool, served as an excellent dataset to test both models on. Not only were the specific stimuli in this dataset never used for training or fitting both models, the method by which the stimuli in this dataset were obtained also differed from the method used to obtain the training stimuli. Presumably, this has resulted in different stimulus characteristics.

For the stimuli generated with this tool, we found that movement trajectory pairs generated by participants who are instructed to pursue individual goals as well as trajectory pairs generated by shuffling other trajectory pairs, receive the lowest social interaction ratings from participants as well as all predictors we studied. This outcome is intuitively plausible, since the movements of each agent in these videos were not intentionally controlled in order to react to the movements of the other visible agent. On the other hand, stimuli in the courting and fighting categories, as well as most stimuli in the chasing category, received high social interaction ratings, presumably because they reflect intentional interaction. Stimuli in the free interaction and hindering categories received a wider variety of social interaction scores, possibly because these behaviours are more complex, and were therefore interpreted or executed in different ways by different participants who created them.

Theoretical implications and explanations of our findings Our finding that a higher proximity (or, alternatively, a smaller distance) between agents results in higher social interaction ratings is consistent with earlier studies (Rasmussen & Jiang, 2019; Roux et al., 2013). An exponential relation seems plausible, however, since it corresponds to intuition: if one wants to judge whether two people are interacting, the difference between them standing 0.5 or 2.5 metres apart (both relatively high proximities) seems to be quite important for this judgement, whereas the difference between them standing 10 or 12 metres apart (both relatively low proximities) seems much less significant. Yet, in both cases the difference in distance, and hence in proximity, is equal.

In contrast to our findings, Gao and Scholl (2011) claim that animacy detection in stimuli containing more than one agent, cannot be based on proximity alone and empirically support this claim. In their *Don't-get-caught* experiments, for example, distractor objects close to a sheep agent where not detected as animate as long as they moved randomly, whereas a distant wolf agent was detected as animate as long as it was chasing the sheep (Gao et al., 2009; Gao & Scholl, 2011; Scholl & Gao, 2013). Crucially, however, the stimuli they used were more complex than our stimuli: they contained more than two agents (e.g. a wolf, a sheep and distractors), which were sometimes shaped triangularly and faced specific directions. For stimuli only displaying the motion trajectories of two agents however, such as the ones used in the current study and the study by Shu et al. (2018), the exponential proximity model explains human performance better than the hierarchical model.

Our results are compatible with both a low-level perceptual and a high-level cognitive view of human social interaction detection. On the one hand, the observation of proximity or distance really is a low-level perceptual process (we automatically see objects as close to or far away from each other and do not have to make an explicit cognitive effort for this) and thus proximity-based social interaction perception could arise early in visual processing. On the other hand, Gao and Scholl (2011) argue that studies in which participants are explicitly asked to provide judgements on perceived animacy can never directly provide evidence for perception, because as soon as participants are asked to rate the animacy they perceive, this rating can or will be subject to high-level cognitive contemplation. The same would hold for social interaction. Humans might not base their judgements on whether they perceive stimuli as containing social interaction, but rather on whether their perception fits their belief of what social interaction should look like (Scholl & Gao, 2013). In theory, participants could consciously use their proximity perception to base their high-level social interaction judgements on.

In order to more clearly illuminate the concept of low-level social interaction perception specifically, we have added a recall task to our experiment aiming to use hit/miss ratio as an implicit measure of social interaction perception. We found weak but significant correlations between social interaction (as measured by explicit ratings and the predictions of both models) and hit/miss ratio in the recall task. These results are in line with earlier research that shows that animate animations are remembered better (van Buren & Scholl, 2017). Furthermore, this result can be interpreted as evidence for the perceptual nature of social interaction detection, because perceptual processes often determine what humans attend to and remember. However, the results of our recall task should be interpreted with caution as explained in the paragraph *limitations*.

The lesser performance of the hierarchical model in the game dataset, is likely a textbook case of overfitting. Overfitting is a phenomenon in which a model is either too flexible or takes into account too many (irrelevant) data features to be generalizable to a wide variety of data (Hawkins, 2004). In such a case, the model learns to classify data resembling the training data unrealistically well, at the cost of its performance on slightly different data. Considering that the hierarchical model contains many more (hidden) parameters than the exponential proximity model (which only contains two variables), the risk of overfitting the hierarchical model is much larger than the risk of overfitting the exponential proximity model. The fact that the hierarchical model fits well on the drone dataset but was found to fit considerably worse on the game dataset (containing fundamentally different stimuli) strongly supports the conclusion that this is a case of overfitting. Moreover, Shu et al. (2018) themselves also found a considerably worse fit when testing their model on data of a different kind than the data they trained their model on. When fitting the exponential proximity model on the drone data, we were not able to match the hierarchical model performance on the same data, but this performance did not drop dramatically when testing the model on the game data. (It even improved in some aspects.) From this we can conclude that the exponential proximity model did not overfit.

Limitations The social interaction detection experiment we conducted did not take place in a controlled lab setting. Instead participants used their own personal computers or mobile devices and could participate from any location they wanted. Because of the Covid-19 pandemic and measures taken against the pandemic, we had no other option than to design an online experiment, but the lack of a controlled lab environment might have influenced our results. In order to have some control over the experimental conditions, we instructed participants to participate from a place free of distraction, to focus on the experiment completely and to only take an optional break between the rating task and judgement task. Furthermore participants had to indicate the extent to which they had been distracted and whether they were interrupted for longer than 10 minutes. As shown in the results excluding distracted and interrupted participants did not heavily influence our main results however.

As pointed out above a social interaction detection task in which participants are asked to judge social interaction ratings can never unambiguously provide evidence for low-level, perceptual processing. Because of that we added a recall task to our study, but since many factors can influence whether an animation is remembered correctly and since the correlations we found between hit/miss ratio and other social interaction measures were rather weak, we cannot use the recall rate as an unambiguous implicit measure of social interaction perception. Future studies which do control more carefully for other stimulus factors influencing memory of stimuli, or use other implicit measures that are associated with perception, are needed to determine whether proximity and the exponential proximity model predict low-level social interaction perception, instead of cognitive social interaction judgements. Based on this study alone, it is safer to use the more neutral term ‘human social interaction detection’.

Finally, it could be argued that the focus of this study is rather limited. There is more to perceptual or cognitive social interaction processing than simply detecting whether two objects are engaging in social interaction. We might respond differently to different categories of social interaction (e.g. fighting versus playing) and in distinguishing these categories proximity might not be a helpful cue, whereas the hierarchical model might be relatively easily adapted by explicitly linking the subinteraction layer to social category labels and thereby be more suitable for explaining more subtle aspects of social interaction processing. On the other hand it could also be possible that a simpler model based on proximity in combination with other low-level stimulus features suffices to explain this. Additionally, as pointed out above, proximity might also not be the most helpful cue for stimuli more complex than animations only showing the motion trajectories of two objects. Hence future research focussing on different stimulus features and models explaining more subtle forms of social interaction processing (e.g. categorizing social interactions) and social interaction processing in more complex stimuli is needed.

6. Conclusion

The present study aimed to determine whether the hierarchical model designed by Shu et al. (2018) based on subinteraction classification and temporal interactivity parsing could explain human social interaction detection in movement trajectories of two objects. We found that this model does not provide a plausible explanation. The average proximity between two agents, especially when used as input to an exponential model, provides an alternative explanation that does not explain experimental findings of human social interaction ratings considerably worse and in some aspects even better. We prefer the simpler alternative explanation over the more complex hierarchical model. Although proximity is a low-level cue, our findings did not unambiguously implicate low-level social interaction detection and therefore does not exclude the possibility of human social interaction detection itself being a high-level cognitive process, building on low-level processes.

Implications for the field of Humanities

Our study provided evidence that at least parts of human social interaction detection can be reduced to rule-following and that the rules that are followed are in this case probably very simple calculations based on proximity. This observation could be the starting point of theories arguing that even something seemingly complex as the rich social behavior of human beings, might ultimately be reducible to rule-following behaviour. This would support a much more mechanistic view on humanity.

Additionally the fact that humans interpret two dots as socially interacting as long as they are close to each other, shows how adept humans are in scanning their environment for potential social agents. This supports a philosophical view of human beings as intrinsically social creatures.

7. Acknowledgements

First and foremost, I would like to express my great gratitude to Dr. Tianmin Shu for giving me access to the model he and his colleagues designed and the data they collected. Without his co-operative attitude and fast responses to our requests, this thesis could not have been written. Next I want to thank my thesis supervisor, Dr. Ruud Hortensius, for his enthusiasm in guiding me in the process of writing a thesis and the time he spent helping me. Finally, I want to thank all of my friends and family members who spent time watching and rating animations of moving dots, testing or playing the trajectory generation game, relieving my stress or otherwise contributing to the creation of this thesis. Special thanks go to my brother Joris, who greatly improved the readability and consistency of this thesis by thoroughly proof-reading this thesis.

8. Literature

Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Do visual perspective tasks need theory of mind? *Neuroimage*, *30*(3), 1059–1068.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*(4), 313–331.

Blythe, P. W. (1999). How motion reveals intention: categorizing social interactions. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 257–286). Oxford University Press, USA.

Chandler-Mather, N., Welsh, T., Sparks, S., & Kritikos, A. (2020). Biological motion and animacy belief induce similar effects on involuntary shifts of attention. *Attention, Perception, & Psychophysics*, *82*(3), 1099–1111. https://doi.org/10.3758/s13414-019-01843-z

Dittrich, W. H., & Lea, S. E. G. (1994). Visual perception of intentional motion. *Perception*, *23*(3), 253–268.

Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 671–678.

Frith, C. D. (2008). Social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1499), 2033–2039.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*(12), 1845–1853.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, *59*(2), 154–179.

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 669.

Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Hovaidi-Ardestani, M., Saini, N., Martinez, A. M., & Giese, M. A. (2018). Neural Model for the Visual Recognition of Animacy and Social Interaction. *International Conference on Artificial Neural Networks*, 168–177.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152.

Isik, L., Mynick, A., Pantazis, D., & Kanwisher, N. (2020). The speed of human social interaction perception. *NeuroImage*, *215*, 116844.

McAleer, P., & Pollick, F. E. (2008). Understanding intention from minimal displays of human activity. *Behavior Research Methods*, *40*(3), 830–839.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Rasmussen, C. E., & Jiang, Y. v. (2019). Judging social interaction in the Heider and Simmel movie. *Quarterly Journal of Experimental Psychology*, *72*(9), 2350–2361.

Roux, P., Passerieux, C., & Ramus, F. (2013). Kinematics matters: A new eye-tracking investigation of animated triangles. *Quarterly Journal of Experimental Psychology*, *66*(2), 229–244. https://doi.org/10.1080/17470218.2012.704052

Scholl, B. J., & Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment. *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention*, *4629*, 197–229.

Schultz, J., & Bülthoff, H. H. (2019). Perceiving animacy purely from visual motion cues involves intraparietal sulcus. *NeuroImage*, *197*, 120–132.

Schultz, J., Friston, K. J., O'Doherty, J., Wolpert, D. M., & Frith, C. D. (2005). Activation in posterior superior temporal sulcus parallels parameter inducing the percept of animacy. *Neuron*, *45*(4), 625–635.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34. https://doi.org/https://doi.org/10.1016/j.neubiorev.2014.01.009

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. (2017). Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: A review using probabilistic atlases from different imaging modalities. *Human Brain Mapping*, *38*(9), 4788–4805. https://doi.org/https://doi.org/10.1002/hbm.23675

Shu, T., Peng, Y., Fan, L., Lu, H., & Zhu, S. (2018). Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. *Topics in Cognitive Science*, *10*(1), 225–241.

Stewart, J. A. (1982). *Perception of animacy*. (Doctoral dissertation, University of Pennsylvania).

Su, J., van Boxtel, J. J. A., & Lu, H. (2016). Social interactions receive priority to conscious perception. *PloS One*, *11*(8), e0160468.

Tremoulet, P. D., & Feldman, J. (2000). Perception of Animacy from the Motion of a Single Object. *Perception*, *29*(8), 943–951. https://doi.org/10.1068/p3101

van Buren, B., & Scholl, B. J. (2017). Minds in motion in memory: Enhanced spatial memory driven by the perceived animacy of simple shapes. *Cognition*, *163*, 87–92.

Walbrin, J., Downing, P., & Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia*, *112*, 31–39.

Appendix A: Supplementary material

A: Position notation

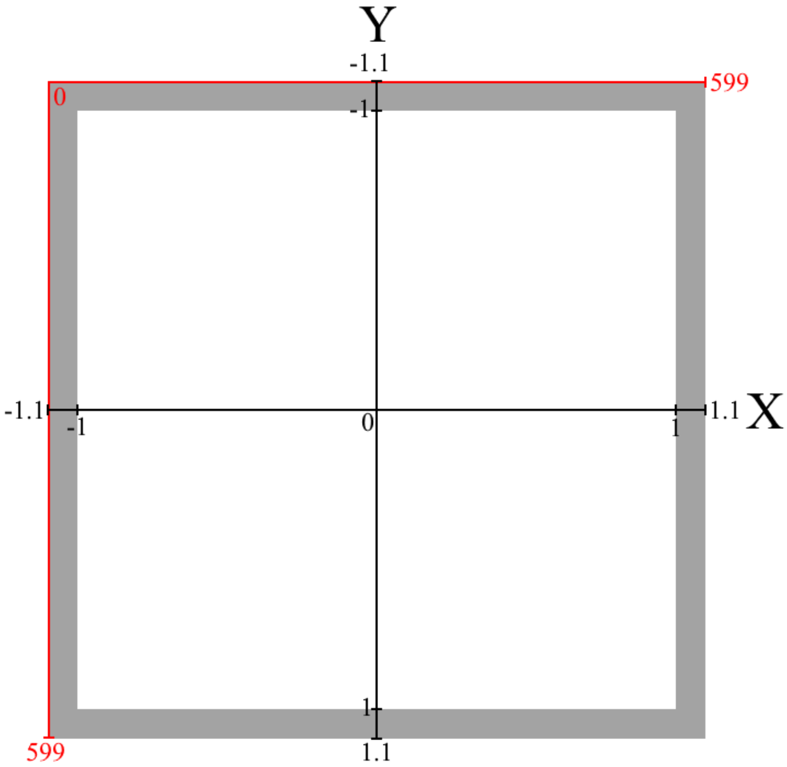


Figure 6: Two ways of notating the positions of agents. In black we pictured the coordinate system used for storing positions in the programs internal state and saving trajectories. Within this system the horizontal and vertical position of both agents was always between -1 and 1. These position reflected the centres of these agents. In red we pictured the indexing system of the 600x600 pixel grid used for visualizing trajectories both real-time during stimulus generation and when generating videos from trajectories. To make sure circles would always be completely visible, frames covered all horizontal and vertical positions always between -1.1 and 1.1, resulting in a border area which was unreachable for the centres of the agents, which we highlighted here in grey.

B: controls

Key for player 1	**Key for player 2**	**When pressed…**	**When released…**
W	Arrow Up	Accelerate in upward direction	Slow down in upward direction
A	Arrow Left	Accelerate in left direction	Slow down in left direction
S	Arrow Down	Accelerate in downward direction	Slow down in downward direction
D	Arrow Right	Accelerate in right direction	Slow down in right direction
Spacebar	Enter	Increase maximal speed	Return maximal speed to normal

Table X1: The controls of both players during stimulus generation.

C: Instructions and settings stimulus generation

Type of behaviour and instructions and information players are provided with	Maximal speed (standard)	Maximal speed (with boost)
Free social interaction: Both players control the circles at their own discretion with the purpose of showing interaction with the other circle.	**Both:** 0.3	**Both:** 0.45
Chasing The chaser chases the other player with the purpose of touching the other player. The other player tries to prevent this by moving away from the chaser. The chased player will have a higher speed limit in order make them more capable of avoiding the chaser.	**Chaser:** 0.30 **Chased:** 0.36	**Chaser:** 0.30 **Chased:** 0.57
Individual goals Two goal circles will appear with colours corresponding to both player colours. The players have to move towards the circle of the corresponding colour, After touching the goal circle of a corresponding colour it will disappear and reappear at another location which the player will have to move to next. The speed at which players reach their goals does not matter, there should be no competition between players and players should ignore each other as much as possible.	**Both:** 0.3	**Both:** 0.45
Hindering Only goal circles in the hindered player's colour appear. The hindered player has the same purpose as in the preceding condition. The hindering player has the purpose of preventing the other player from reaching their goal. The hindering player can do this by getting in the way of the other player. In this condition the hinderer has a higher speed boost to help them hindering the other player.	**Both:** 0.3	**Hinderer:** 0.495 **Hindered:** 0.375
Fighting Both players show fighting behaviour by (literary) bumping into each other.	**Both:** 0.3	**Both:** 0.45
Courting The courter moves in a way they think will catch the attention and interest of the other player. The other player tries to show interest or disinterest in their movements. Both players move slower than normal.	**Both:** 0.225	**Both:** 0.375

Table 4: An overview of the types of social interaction used for the generation of stimuli with instructions and information provided to players and adjusted settings. Speed is measured in distance between trajectory coordinates per second.

D: The difference in $\overline{\text{RMSE}}$ for different predictors on both datasets.

	Dataset	Proximity	Exponential proximity model	Hierarchical model
Hierarchical model	Game dataset	−.092	− .014	
	Drone dataset	− .19	− .027	
Exponential proximity model	Game dataset	− .078		+ .014
	Drone dataset	− .16		+ 0.027

Table 5: The difference in the mean RMSE for all predictors. The numeric values in each cell represent the difference between the $\overline{\text{RMSE}}$ resulting from using the row predictor and the $\overline{\text{RMSE}}$ resulting from using the column predictor. The second column specifies the dataset used to obtain both $\bar{r}$'s. All original $\overline{\text{RMSE}}$ -values can be found in table 2. Green cells indicate a significant negative difference ($p < .01$), orange cells indicate no significant difference and red cells indicate a significant positive difference ($p < .01$). To determine significance we used paired two-sampled t-tests. See supplementary material E for the corresponding t-values.

E: t-values for the difference between samples of Pearson corelations between different predictors and individual participant data.

A: t-values for the individual Pearson corelation coefficients.

	Dataset	Proximity	Exponential proximity model	Hierarchical model
Hierarchical model	Game dataset	$t(32) = -1.0365$	$t(32) = -3.8191$	
	Drone dataset	$t(32) = 7.9954$	$t(32) = 5.9449$	
Exponential proximity model	Game dataset	$t(32) = 3.6611$		$t(32) = 3.8191$
	Drone dataset	$t(32) = 4.0249$		$t(32) = -5.9449$

B: t-values for the individual RMSE-values.

	Dataset	Proximity	Exponential proximity model	Hierarchical model
Hierarchical model	Game dataset	$t(32) = -6.9627$	$t(32) = -.87699$	
	Drone dataset	$t(32) = -31.282$	$t(32) = -3.4944$	
Exponential proximity model	Game dataset	$t(32) = -2.8417$		$t(32) = .87699$
	Drone dataset	$t(32) = -15.984$		$t(32) = 3.4944$

Table 6A: The t-values in the cells are the result of a paired two sample t-test on the difference between the sample of Pearson correlation coefficients between the predictions from the *row* predictor and all individual social interaction ratings and the sample of Pearson correlation coefficients between the predictions from the *column* predictor and all individual social interaction ratings. The second column specifies the experimental data used to find the correlations. **B:** This table is equal to table 6A except that RMSE-values were used instead of Pearson corelation coefficients.

F: Hierarchical model predictions versus experimental rating excluding distracted and interrupted participants

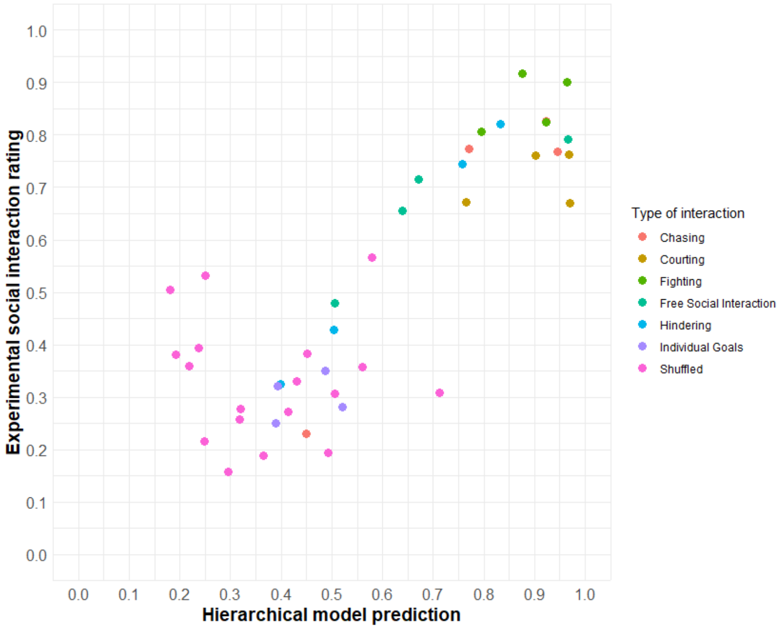


Figure 7: The predictions of the hierarchical model by Shu et al. (2018) versus the mean of the explicit experimental social interaction ratings of all participants in the game dataset. Each dot represents an animation. The colour of each dot shows the interaction label of this animation. Distracted and interrupted participants are excluded.

Appendix B: Incorporation of the Humanities Honours Programme's (HHP) learning pathways

This thesis incorporates the HHP learning pathways of academic depth and interdisciplinary breadth. The fact this thesis is not simply limited to testing the hierarchical model on novel stimuli, but also included generating these stimuli ourselves, including hit/miss ratio in a recall task as a possible implicit social interaction measure and most importantly a comparison with proximity and an exponential proximity model (which I created and fitted myself) is strong evidence for the academic depth of this study. I was not satisfied when I found that the hierarchical model performed well on the game dataset, but wanted to make sure that there were no simpler, alternative explanations for human social interaction detection. This thesis shows interdisciplinary breadth in the combination of literature and methods from cognitive psychology (especially recognizable in the social interaction detection experiment) and computer sciences (especially recognizable in the trajectory generation program and computational models.)