frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 09 June 2021
doi: 10.3389/fnbot.2021.648527

A Dynamical Generative Model of Social Interactions

*Alessandro Salatiello†, Mohammad Hovaidi-Ardestani† and Martin A. Giese**

Section for Computational Sensomotorics, Department of Cognitive Neurology, Centre for Integrative Neuroscience, Hertie Institute for Clinical Brain Research, University Clinic Tübingen, Tübingen, Germany

The ability to make accurate social inferences makes humans able to navigate and act in their social environment effortlessly. Converging evidence shows that motion is one of the most informative cues in shaping the perception of social interactions. However, the scarcity of parameterized generative models for the generation of highly-controlled stimuli has slowed down both the identification of the most critical motion features and the understanding of the computational mechanisms underlying their extraction and processing from rich visual inputs. In this work, we introduce a novel generative model for the automatic generation of an arbitrarily large number of videos of socially interacting agents for comprehensive studies of social perception. The proposed framework, validated with three psychophysical experiments, allows generating as many as 15 distinct interaction classes. The model builds on classical dynamical system models of biological navigation and is able to generate visual stimuli that are parametrically controlled and representative of a heterogeneous set of social interaction classes. The proposed method represents thus an important tool for experiments aimed at unveiling the computational mechanisms mediating the perception of social interactions. The ability to generate highly-controlled stimuli makes the model valuable not only to conduct behavioral and neuroimaging studies, but also to develop and validate neural models of social inference, and machine vision systems for the automatic recognition of social interactions. In fact, contrasting human and model responses to a heterogeneous set of highly-controlled stimuli can help to identify critical computational steps in the processing of social interaction stimuli.

Keywords: social interactions, generative model, motion cues, social perception, social inference

OPEN ACCESS

Edited by:
Letizia Marchegiani, Aalborg University, Denmark

Reviewed by:
Ashley Liddiard, Ford Motor Company, United States
Bin Zhi Li, Chongqing Institute of Green and Intelligent Technology (CAS), China

***Correspondence:**
Martin A. Giese
martin.giese@uni-tuebingen.de

†These authors have contributed equally to this work

Received: *31 December 2020*
Accepted: *23 April 2021*
Published: *09 June 2021*

Citation:
Salatiello A, Hovaidi-Ardestani M and Giese MA (2021) A Dynamical Generative Model of Social Interactions. Front. Neurorobot. 15:648527. doi: 10.3389/fnbot.2021.648527

1. INTRODUCTION

Human and non-human primates are able to recognize the social interactions taking place in their environment quickly and effortlessly: with a few glances out of the window, we can easily understand whether two people are following each other, avoiding each other, fighting, or are engaging in some other form of social behavior. Notably, such interactive behaviors can be recognized even when the available visual information is poor: for example, when the scene we are watching is unfolding behind the leaves of a tree, at a considerable distance from us, or in a low-resolution video. In some of these situations, critical visual cues such as facial expressions might be completely occluded, yet our ability to make social inference is largely unaffected. Such perceptual ability is instrumental in allowing us to move in our social environment and flexibly interact with it, while abiding by the social norms (Troje et al., 2013). Therefore, it constitutes an important social skill that is worth characterizing and modeling also for the development of social robots.

Understanding the neural mechanisms underlying the inference of animacy and social interactions from visual inputs is a long-standing research challenge (Heider and Simmel, 1944; Michotte, 1946; Scholl and Tremoulet, 2000; Troje et al., 2013). Recent work has started identifying some of the responsible neural circuits (Castelli et al., 2000; Isik et al., 2017; Sliwa and Freiwald, 2017; Walbrin et al., 2018; Freiwald, 2020). Even though the detailed computational mechanisms mediating the formation of social percepts from visual inputs remain largely unknown, converging evidence has shown that the observation of biological motion alone is enough for humans to make accurate social inferences (e.g., Heider and Simmel, 1944; Tremoulet and Feldman, 2000; McAleer and Pollick, 2008; Roether et al., 2009). For example, Heider and Simmel (1944) demonstrated that humans can reliably decode animacy and social interactions from strongly impoverished stimuli consisting of simple geometrical figures moving around in the two-dimensional plane. Remarkably, despite their highly abstract nature, the visual stimuli used in this study were perceived as *alive* and sometimes even *anthropomorphic*: the agents were often considered as endowed with intentions, emotions, and even personality traits.

Several subsequent studies (e.g., Oatley and Yuill, 1985; Rimé et al., 1985; Springer et al., 1996; Castelli et al., 2000, 2002) replicated these findings using similar stimuli and showed that the inference of social interactions from impoverished stimuli is a cross-cultural phenomenon (Rimé et al., 1985) that is present even in 5-year-old preschoolers (Springer et al., 1996). Taken together, these findings support the view that the perception of animacy and social interactions might rely on some innate and automatic processing of low-level kinematic features present in the visual inputs, rather than on higher-level cognitive processing (Scholl and Gao, 2013).

The identification of the most critical visual features that shape these social percepts has also received great attention (Tremoulet and Feldman, 2000, 2006). For example, influential work suggested that these percepts are mediated by the detection of apparent violations of the principle of conservation of energy (Dittrich and Lea, 1994; Gelman et al., 1995; Csibra, 2008; Kaduk et al., 2013). Later research proved that also agent's orientation, velocity, and acceleration play a major role (Szego and Rutherford, 2008; Träuble et al., 2014). At the same time, neuroimaging work has shed light on some of the brain regions mediating these phenomena: the right posterior superior temporal sulcus (pSTS—Isik et al., 2017; Walbrin et al., 2018), the medial prefrontal cortex (mPFC—Castelli et al., 2000; Sliwa and Freiwald, 2017), and the right temporoparietal junction (TPJ—Castelli et al., 2000; Saxe and Kanwisher, 2003) are among the brain regions most frequently reported as being involved in the perception of social interaction. Interestingly, Schultz and Bülthoff (2019), recently identified another region—the right intraparietal sulcus (IPS)—that seems to be exclusively engaged during the perception of animacy.

Clearly, the success of both behavioral and neuroimaging social perception studies is tightly linked to the ability to finely control the visual stimuli that participants are exposed to. Specifically, such stimuli should ideally be generated through a process that allows complete parametric control, the creation of a high number of replicates with sufficient variety, and the gradual reduction of complexity. *Parametric control* (e.g., over agents' speed) facilitates the identification of brain regions and individual neurons whose activation covaries with the kinematic features of agents' behavior. *Variety* in classes of social interaction allows the characterization of the class-specific and general response properties of such brain regions. *Numerosity* allows averaging out response properties that are independent of social interaction processing. Finally, the ability to control stimulus complexity allows the generation of *impoverished stimuli* that are fundamental to minimize the impact of confounding factors, inevitably present, for example, in real videos. Similarly, such properties are also desirable when designing and validating neural and mechanistic models of human social perceptions: contrasting human and model responses to a variety of highly controlled stimuli can help discriminate between the computational mechanisms that the models capture well from those that need further refinement. This is especially critical for state-of-the-art deep learning models (e.g., Yamins et al., 2014), which can easily have millions of parameters and be prone to over-fitting.

Currently, no well-established method can generate visual stimuli for the analysis of social perception that satisfy all of the above conditions. Because of this, researchers often have to resort to time-consuming and class-specific, heuristic procedures. A creative approach to this problem has been the one adopted by Gordon and Roemmele (2014), where the task of generating videos was assigned to a set of participants—who were asked to create their own videos of socially interacting geometrical shapes, and to label them accordingly. However, typically, researches use visual stimuli where agents' trajectories are hand-crafted or hard-coded (e.g., Heider and Simmel, 1944; Oatley and Yuill, 1985; Rimé et al., 1985; Springer et al., 1996; Castelli et al., 2000, 2002; Baker et al., 2009; Gao et al., 2009, 2010; Kaduk et al., 2013; Träuble et al., 2014; Isik et al., 2017; van Buren et al., 2017; Walbrin et al., 2018), based on rules (e.g., Kerr and Cohen, 2010; Pantelis et al., 2014), or derived from real videos (e.g., McAleer and Pollick, 2008; McAleer et al., 2011; Thurman and Lu, 2014; Sliwa and Freiwald, 2017; Shu et al., 2018). All of these approaches suffer from significant limitations. Hand-crafted trajectories need to be generated *de novo* for each experimental condition and are not easily amenable to parametric control. Likewise, the extraction of trajectories from real videos also comes with its burdens: real videos need to be recorded, labeled, and heavily processed to remove unwanted background information. Rule-based approaches offer an interesting alternative. However, it is generally difficult to define natural classes of social interactions using rules akin to those used in Kerr and Cohen (2010) and Pantelis et al. (2014). Recent work (Schultz and Bülthoff, 2019; Shu et al., 2019, 2020) has generated visual stimuli using model-based methods; however, these models can only generate limited and generic classes of social interaction (namely, cooperative and obstructive behaviors). Finally, specialized literature on the collective behavior of humans and animals has produced a wealth of influential models (Blackwell, 1997; Paris et al., 2007; Luo et al., 2008; Russell et al., 2017); however, such models can

also typically account only for simple behaviors (e.g., feeding, resting, and traveling) and for basic interactions (e.g., avoidance and following).

To overcome the limitations of the above methods, in this work, we introduce a dynamical generative model of social interactions. In stark contrast to previous work, our model is able to automatically generate an arbitrary number of parameterized motion trajectories to animate virtual agents with 15 distinct interactive motion styles; the modeled trajectories include the six fundamental interaction categories frequently used in psychophysical experiments (i.e., *Chasing*, *Fighting*, *Flirting*, *Following*, *Guarding*, and *Playing*—Blythe et al. 1999; Barrett et al. 2005; McAleer and Pollick 2008) and nine relevant others. The model controls *speed*, and *motion direction*, arguably the two most critical determinants of social interaction perception (Tremoulet and Feldman, 2000; Szego and Rutherford, 2008; Träuble et al., 2014). Finally, we validated the model with three psychophysical experiments, which demonstrate that participants are able to consistently attribute the intended interaction classes to the animations generated with our model.

The rest of the paper is organized as follows. In section 2, we describe the generative model and the experiments we conducted to validate it. Next, in section 3, we summarize the experimental results. Finally, in section 4, we (1) explain how our results validate the developed model, (2) explain how the model compares to related work, and (3) discuss the main limitations of our model and future directions.

2. METHODS

2.1. Related Modeling Work

The generative model we introduce in this work builds on classical models of biological and robotic navigation. In the classical work by Reichardt and Poggio (1976), the authors proposed a dynamical model to describe the navigation behavior of flies intent on chasing moving targets as part of their mating behavior. The core idea was to consider the moving targets as *attractors* of the dynamical system describing the flies' trajectories. Subsequently, Schöner and Dose (1992) and Schöner et al. (1995) used a similar approach to develop a biomimetic control system for the navigation of autonomous robots. Critically, such a system was also able to deal with the presence of obstacles in the environment, which were modeled as *repellors*. Extending this system, Fajen and Warren (2003) built a model of human navigation that was able to closely capture the trajectories described by their participants as they walked naturally toward targets while avoiding obstacles on their way. Specifically, this model was able to describe the dynamics of the participants' average heading direction very accurately; however, their speed was roughly approximated as constant.

Alternative approaches can characterize richer navigation behaviors by jointly modeling both heading direction and speed dynamics. This idea was successfully used to control the motion of both autonomous vehicles (Bicho and Schöner, 1997; Bicho et al., 2000) and robotic arms (Reimann et al., 2011). Similar approaches have also been used in computer graphics to model the navigation of articulated agents (Mukovskiy et al., 2013).

2.2. The Generative Model

To model the interactive behavior of two virtual agents, we define, for each agent i, a dynamical system of two nonlinear differential equations. Specifically, the equations describe the dynamics of the agent's heading direction $\phi_i(t)$ and instantaneous propagation speed $s_i(t)$.

The heading direction dynamics, derived from Fajen and Warren (2003), are defined by:

$$\ddot{\phi}_i(t) = -b\dot{\phi}_i(t) + A(\phi_i(t), \psi_i^g(t)) + R(\phi_i(t), \boldsymbol{\psi}_i^o(t)) \quad (1)$$

In this equation, $A(\phi_i(t), \psi_i^g(t))$ defines the *attraction* of agent i to the goal g located along the direction $\psi_i^g(t)$, at a distance $d_i^g(t)$ from it. Similarly, $R(\phi_i(t), \boldsymbol{\psi}_i^o(t))$ defines the *repulsion* of agent i for the obstacles $\boldsymbol{o} = [o_1, o_2, ..., o_{N_{obst}}]^T$ located along the directions $\boldsymbol{\psi}_i^o(t)$, at a distance $\boldsymbol{d}_i^o(t)$ from it. These two functions are given by:

$$\begin{aligned} A(\phi_i(t), \psi_i^g(t)) &= -k^g(\phi_i(t) - \psi_i^g(t))(e^{-c_1 d_i^g(t)} + c_2) \\ R(\phi_i(t), \boldsymbol{\psi}_i^o(t)) &= k^o \sum_{n=1}^{N_{obst}} r^{o_n}(\phi_i(t)) \end{aligned} \quad (2)$$

The contributions of the individual obstacles to the repulsion function are given by:

$$r^{o_n}(\phi_i(t)) = (\phi_i(t) - \psi_i^{o_n}(t))(e^{-c_3|\phi_i(t) - \psi_i^{o_n}(t)|})(e^{-c_4 d_i^{o_n}(t)}) \quad (3)$$

In these equations, k^j and c_j are constants; o_n indicates the *nth* obstacle. Note that, in general, $\psi_i^{o_n}(t)$, which is the direction of the *nth* obstacle of the *ith* agent is time-dependent; for example, depending on the specific social interaction class it might be a function of the instantaneous heading direction of other agents.

The propagation speed dynamics are specified by the following stochastic differential equation:

$$\tau \dot{s}_i(t) = -s_i(t) + F_i(d_i^g(t)) + k_i^\epsilon \epsilon_i(t) \quad (4)$$

where $\epsilon_i(t)$ is Gaussian white noise. The nonlinear function F_i specifies how the agent's speed changes as a function of the distance from its goal:

$$F_i(d) = \frac{c_5}{1 + e^{-c_6^i(d - c_7^i)}} - c_8^i e^{-k_i d} + c_9^i \quad (5)$$

Critically, we choose this specific functional form because it provides us with enough flexibility to reproduce several relevant interaction classes, including the six fundamental interaction categories traditionally studied in psychophysical experiments (Blythe et al., 1999; Barrett et al., 2005; McAleer and Pollick, 2008): *Chasing*, *Fighting*, *Flirting*, *Following*, *Guarding*, and *Playing*.

To generate the trajectories, we first randomly sample a series of goal points for the first agent from a two-dimensional uniform distribution over the 2D plane of action. Such goal points are commonly referred to as *via points*. We then use the instantaneous position of the first agent as goal position for the

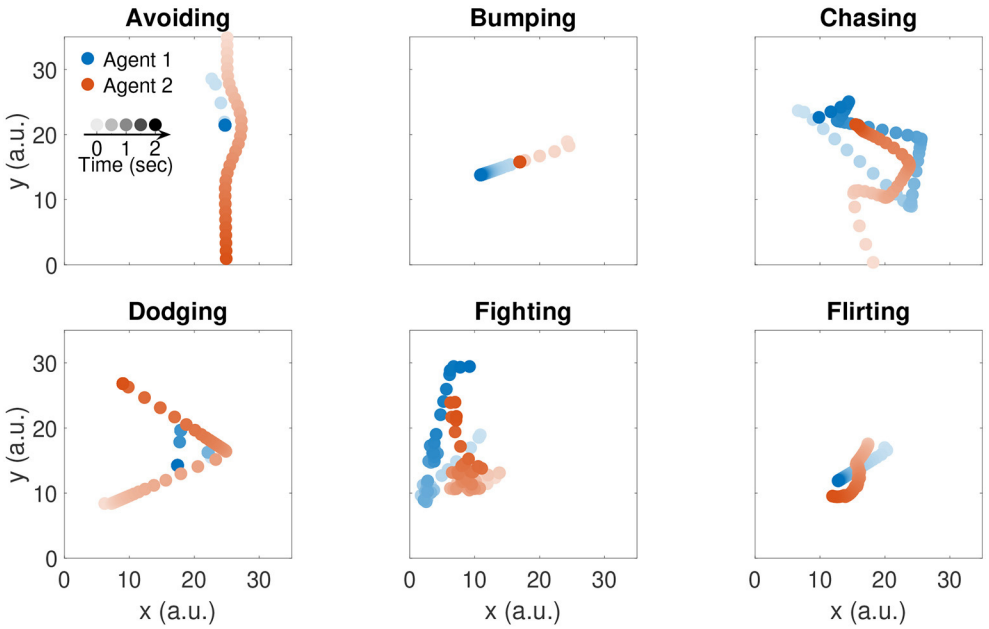

FIGURE 1 | Trajectories of six example social interactions. Color indicates agent identity: agent 1 is represented in blue; agent 2 is represented in red. Color saturation indicates time: darker colors indicate recent time samples.

Algorithm 1: Pseudocode for trajectory generation

Input: Class-specific parameters θ_c
Output: Agents' direction Φ and speed S
for each timestep t **do**
  for each agent i **do**
    compute goal direction $\psi_i^g(t)$
    compute distance from goal $d_i^g(t)$
    for each obstacle o_n **do**
      compute obstacle direction $\psi_i^{o_n}(t)$
      compute distance from obstacle $d_i^{o_n}(t)$
    end
    compute $\phi_i(t)$ integrating Equation (1)
    compute $s_i(t)$ integrating Equation (4)
  end
end
/*Note: $\psi_i^g(t)$ and $\psi_i^{o_n}(t)$ are either specified a priori or computed dynamically depending on the agent and social interaction class. For example, for simple behaviors (e.g., chasing) $\psi_1^g(t)$ and $\psi_1^{o_n}(t)$ are specified a priori, while $\psi_2^g(t) = \phi_1(t)$*/

second agent. Samples that are too close to the current agent's position are rejected. Further details about the implementation of the generative model are provided in the Algorithm 1 box. Representative trajectories of six example social interactions are illustrated in **Figure 1**. Note that the speed control dynamics are not influenced by the presence of obstacles, since their effect was not needed to realistically capture the social interactive behaviors we chose to model.

2.3. Model Validation

To assess whether our model is able to generate perceptually valid socially interactive behaviors, we carried out three behavioral experiments. In these experiments, we asked participants to categorize videos of interacting agents generated with our model in a free-choice task (Experiment 1), and in a forced-choice task (Experiment 2). Finally, we analyzed the semantic similarities between the labels chosen by the participants (Experiment 3).

2.3.1. Dataset Generation

To validate our approach, we chose to model the six fundamental interaction classes (i.e., *Chasing*, *Fighting*, *Flirting*, *Following*, *Guarding*, and *Playing*; Blythe et al. 1999; Barrett et al. 2005; McAleer and Pollick 2008), and nine other relevant ones (i.e., *Avoiding*, *Bumping*, *Dodging*, *Frightening*, *Meeting*, *Pulling Pushing*, *Tug of War*, and *Walking*) resulting in a total of 15 interaction classes. To generate the trajectories corresponding to these classes, we simulated the model with 15 distinct parameter sets, which we identified through a simulation-based heuristic procedure. A list of the most critical parameters is presented in **Table 1**. The complete dataset we generated for our experiments included five random realizations of each interaction class, for a total of 75 videos. Each random realization is defined by different via points and noise realizations.

2.3.2. Participants

A total of 39 participants with normal or corrected vision took part in the experiments: 13 in Experiment 1 (9 females, 4 males), ten in Experiment 2 (5 females, 5 males), and 16 in Experiment 3 (9 females, 7 males). All participants were college students attending the University of Tübingen and provided written

TABLE 1 | Main model parameters.

Interaction class	Agent 1							Agent 2						
	k	k^{ϵ}	c_5	c_6	c_7	c_8	c_9	k	k^{ϵ}	c_5	c_6	c_7	c_8	c_9
Avoiding	0	0	1	1	5	3	0	0	0	0.4	1	0	2.7	0
Bumping	0	0.9	1	0.8	0	0	0	0	1	0.8	10	0	1	0
Chasing	0	0	1	10	7	0	0	0	0	1	1	7	0	0
Dodging	0	0	1	0.5	7	5	0	0	0	3	1	0	0	0
Fighting	0.1	0	1	1	3	1	0	0.1	1	1	1	3	1	0
Flirting	0	0	1	1	5	0	0	0.5	1	0.6	1	2	1	0
Following	0	0	1	10	7	0	0	0	0	1	4	4	0	0
Frightening	0	0	1	1	5	0	0	0	0	1	1	5	0	0.5
Guarding	0	0	1	1	5	0	0	0	0	1	1	3	0	0.5
Meeting	0	0.2	1	2	0	6	0	0.5	1	0.22	3	0	6	0
Playing	0	0	1	1	5	0	0	0	1	1	1	10	0	0.5
Pulling	0	0	1	10	0	2.6	0	0	0	0.9	5	0	2.6	0
Pushing	0	0	1	10	0	2.5	0	0	0	0.1	1	0	0	2.5
Tug of War	0	0.2	1	10	0	6	0	0	0.5	0.9	5	0	0	0.5
Walking	0	0.2	1	10	0	1	0	0	0	0.22	10	0	0	0

informed consent before the experiments. All experiments were in full compliance with the Declaration of Helsinki. Participants were naïve to the purpose of the study and were financially compensated for their participation.

2.3.3. Experiment Setup

In Experiment 1 and Experiment 2, participants sat in a dimly lit room in front of an LCD monitor (resolution: 1,920 × 1,080, refresh rate: 60Hz), at a distance of 60cm from it. To ensure that all participants would observe the stimuli with the same view parameters and the same distance from the screen, they were asked to place their heads in a chin-and-forehead rest during the experimental sessions. The experiments started with a short familiarization session during which the participants learned to use the computer interface. Subsequently, the participants were shown the videos generated with our model. Their task was to describe the videos by using their own words (Experiment 1) or by selecting labels among those provided to them (Experiment 2), and to provide animacy ratings through a standard 0–10 Likert scale. To increase the confidence in their answers, we gave participants the opportunity to re-watch each video up to three times. The videos were presented in pseudo-randomized order over five blocks. Five-minute rest breaks were given after each block. The animated videos always showed two agents moving in a 2D plane following speed and direction dynamics generated offline with our model. Critically, unlike in previous work (Blythe et al., 1999; Barrett et al., 2005), our agents were very simple geometrical shapes, namely a blue circle and a red rectangle (as in Tremoulet and Feldman, 2000); this choice ensured that participants' perception would not be biased by additional visual cues beyond the agents' motion and relative positions. In Experiment 3, subjects were asked to fill out a questionnaire to rate the semantic similarity between social interaction classes (0–10 Likert scale).

2.3.4. Experiment 1

The first experiment was aimed at assessing whether subjects would perceive the motion of virtual agents generated with our model as a social interaction. The second goal of this experiment was the identification of unequivocal labels for the interaction classes generated with our model. To this end, we asked participants to watch all the videos in our stimulus set (section 2.3.1). After watching the videos, subjects were asked to provide their own interpretations by summarizing what they had perceived with a few sentences or keywords. Importantly, in this experiment, to make sure we would not bias the participants' perceptions, we did not provide them with any labels or other cues: they had to come up with their own words. In addition, subjects were asked to provide an animacy rating for each agent. The most commonly reported keywords were used as *ground-truth* interaction labels for the remaining experiments.

To test whether participants assigned different animacy ratings depending on agent identity and social interaction class, we fitted a linear mixed-effect model to the animacy ratings, with Agent and Social Interaction as fixed effects, and Subject as random effect:

$$\begin{aligned}\text{Animacy}_{sl} = \alpha_0 &+ \sum_{i=1}^{N_a} \beta_i \cdot \text{Agent}(i,l) \\ &+ \sum_{i=1}^{N_c} \gamma_i \cdot \text{SocialInteraction}(i,l) + b_{0s} + \epsilon_{sl} \quad (6)\end{aligned}$$

In this model, Animacy_{sl} is the lth animacy rating reported by subject s, with $s = 1, 2, ..., N_s$ and $l = 1, 2, ..., N_a N_c$; N_a, N_c, and N_s are the number of agents, social interaction classes, and subjects, respectively. Moreover, $\text{Agent}(i,l)$ is a dummy variable that is equal to 1 when the rating l is for agent i, and 0 otherwise. Similarly, $\text{SocialInteraction}(i,l)$ is a dummy variable that is equal

to 1 when the rating l is for social interaction i, and 0 otherwise. Finally, b_{0s} is the subject-specific random effect [$b_{0s} \sim N(0, \sigma_b^2)$] and ϵ_{sl} are the residual error terms [$\epsilon_{sl} \sim N(0, \sigma^2)$]. Notably, the model was fitted with a sum-to-zero constrain, that is $\sum_{i=1}^{N_a} \beta_i = 0$ and $\sum_{i=1}^{N_c} \gamma_i = 0$; therefore, in this model, α_0 represents the overall average animacy rating. All the analyses described in this and in the next sections were performed in MATLAB R2020a (The MathWorks, Natick, MA).

2.3.5. Experiment 2

The second experiment was aimed at further studying the social interaction classes perceived by the participants while watching our animated videos. To this end, new subjects were exposed to a subset of the videos in our original dataset. Specifically, for this experiment we excluded the videos corresponding to the classes *Following*, *Guarding*, and *Playing*, as these tended either to be often confused with other classes, or to be labeled with a broad variety of related terms. Critically, unlike in Experiment 1, after watching the videos, participants were asked to describe the videos by choosing up to three labels, among those selected in Experiment 1.

To assess the classification performance, we computed the confusion matrix M. In this matrix, each element $m_{i,j}$ is the number of times participants assigned the class j to a video from class i. Starting from M, we computed, for each social interaction class, Recall, Precision, and F_1 score. Recall measures the fraction of videos of class i that are correctly classified, and is defined as $Recall_i = m_{i=j} / \sum_{j=1}^{N_c} m_{i,j}$. Precision measures the fraction of times participants correctly assigned the class j to a video, and is defined as $Precision_j = m_{i=j} / \sum_{i=1}^{N_c} m_{i,j}$. Finally, the F_1 score is the harmonic mean of Precision and Recall; it measures the overall classification accuracy and is defined as $F_1 = 2 \cdot Precision \cdot Recall / (Precision + Recall)$.

To evaluate whether some classes were more likely to be confused with each other, we computed, for each pair of classes (i, j), with $i \neq j$, the empirical pairwise mislabeling probability, defined as $P_{MS}(i, j) = (m_{i,j} + m_{j,i}) / (\sum_{k=1}^{N_c} \sum_{l \neq k} m_{k,l})$.

To assess whether participants improved their classification performance during the experiment, we computed the average Precision, Recall, and F_1 score across social interaction class, as a function of experimental block; we then fitted linear models to test whether experimental block explained a significant fraction of variation in the performance measures defined above.

2.3.6. Experiment 3

The third and last experiment was aimed at assessing whether there are interpretable semantic similarities among the labels provided in Experiment 2. Some interaction classes were misclassified by the participants in Experiment 2. This suggests that either the generated animated videos are not distinctive enough or that the classes semantically overlap with each other. To disambiguate between the two options, we ran a semantic survey test with a new set of participants. Participants in this experiment did not watch any video. After providing them with precise definitions for each social interaction class, we asked them to indicate the level of semantic similarity for each pair of classes, by providing rates ranging from 0 to 10. Specifically, using this scoring system, participants were asked to assign 0 to pairs of classes perceived as not sharing any semantic similarity, and 10 to those perceived as equivalent classes.

To assess the geometry of the semantic similarity space, we first transformed all the similarity ratings s into distance ratings d by computing their complement (i.e., $d = 10 - s$), and then rescaled them between 0 and 1. All the resulting semantic distances collected from participant i were then stored in a matrix D^i. In this matrix, $D^i_{j,k} = 0$ if the classes j and k were considered as semantically equivalent by subject i; $D^i_{j,k} = 1$ if the classes j and k were considered as semantically unrelated. We then used non-metric multidimensional scaling (MDS; Shepard, 1962a,b) to visualize in a 2D space the underlying relational structure contained in the distance matrix.

To determine whether some groups of classes were consistently considered as semantically similar, we performed agglomerative hierarchical clustering on the distance matrix D using the Ward's linkage method (Ward, 1963), which minimizes the within-cluster variance. Clusters were then identified using a simple cut-off method, using as a threshold $\tau = 0.7 \cdot M_{WD}$, where M_{WD} is the maximum observed Ward's distance.

Finally, to estimate whether the semantic similarity between pairs of classes explained the mislabelings observed in Experiment 2, we computed the Pearson's correlation coefficient (ρ) between the empirical mislabeling probability $P_{MS}(j, k)$ measured in Experiment 2 and the semantic distance $D(j, k)$.

3. RESULTS

3.1. Experiment 1

As mentioned above, participants in this experiment were completely free to provide interpretations about the videos through either labels or short sentences. For each video class, we pooled together all the definitions and labels, and we considered the most used term as the *ground-truth* class label. **Figure 2** summarizes the reported labels for six example social interaction classes. The pie charts show that some classes such as *Avoiding* and *Fighting* tended to be consistently described with very few labels (i.e., 2 − 3). Other classes such as *Dodging* were instead described with more labels (i.e., 6). Regardless of the number of labels used to describe a social interaction class, these were generally semantically similar. For example, some classes were named interchangeably depending on the perspective from which subjects reported their interpretation about the videos. A typical example of this issue is the ambiguity between the classes *Pulling* and *Pushing*. On the other hand, some other classes (for instance *Bumping* and *Pushing*) were sometimes misclassified regardless of the perspective from which subjects might have observed the videos.

Average animacy ratings are reported in **Figure 3A**, with classes sorted in ascending order of average across-agent animacy. Agents, were consistently perceived as animate [$\alpha_0 = 53.27\%, t_{(299)} = 11.72, p = 2.3 \cdot 10^{-26}$]. This is consistent with the fact that self-propulsion (Csibra, 2008), goal directedness (van Buren et al., 2016), being reactive to social contingencies (Dittrich

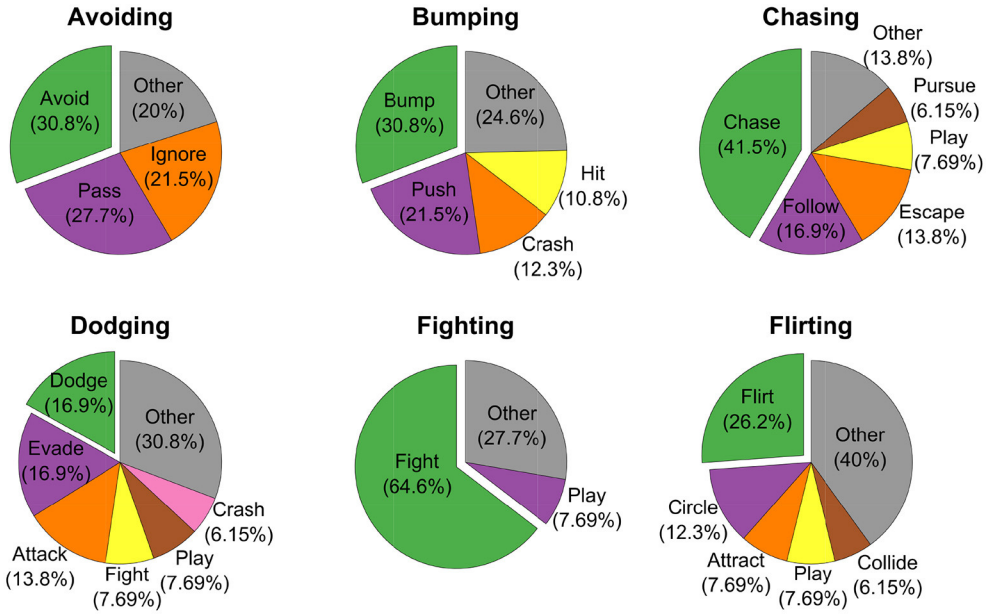

FIGURE 2 | Distribution of reported keywords for six example social interactions. Pie charts' titles indicate the true classes. Individual slices are assigned to all the keywords reported in Experiment 1 occurring with a frequency >5%. Keywords reported with a frequency <5% are pooled together in the slice *Other* (in gray). Offset slices (in green) represent the most frequently reported keywords.

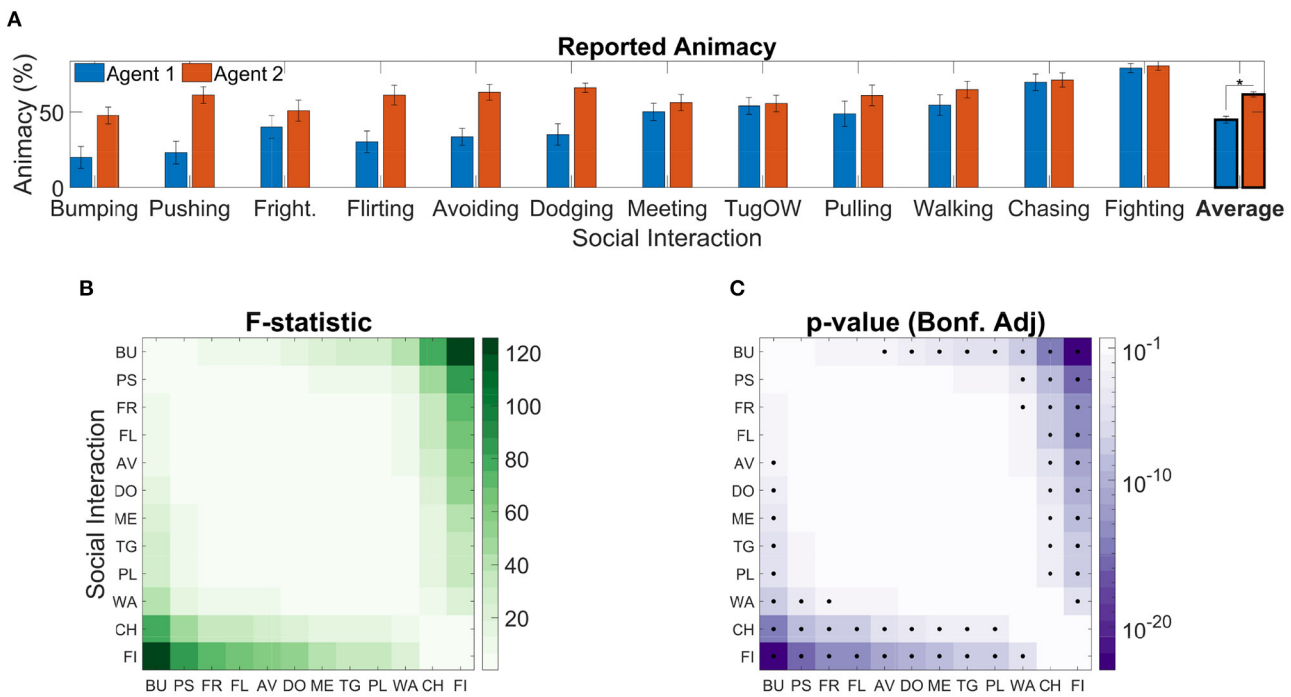

FIGURE 3 | Reported agent animacy. **(A)** Mean animacy ratings obtained in Experiment 1; error bars represent standard errors; results are rescaled between 0 and 100. Classes are sorted in ascending order by average across-agent animacy rating. The asterisk denotes a significant effect ($p < 0.05$) of Agent on Animacy [$F_{(1,299)} = 99.98$, $p = 1.74 \cdot 10^{-20}$]. **(B)** F-statistics of *post-hoc* tests to assess the difference in animacy ratings between social interaction classes [i.e., $F_{(1,299)}$]. **(C)** Bonferroni adjusted *p*-values corresponding to the F-statistics reported in **(B)**; black dots represent significant pairwise differences.

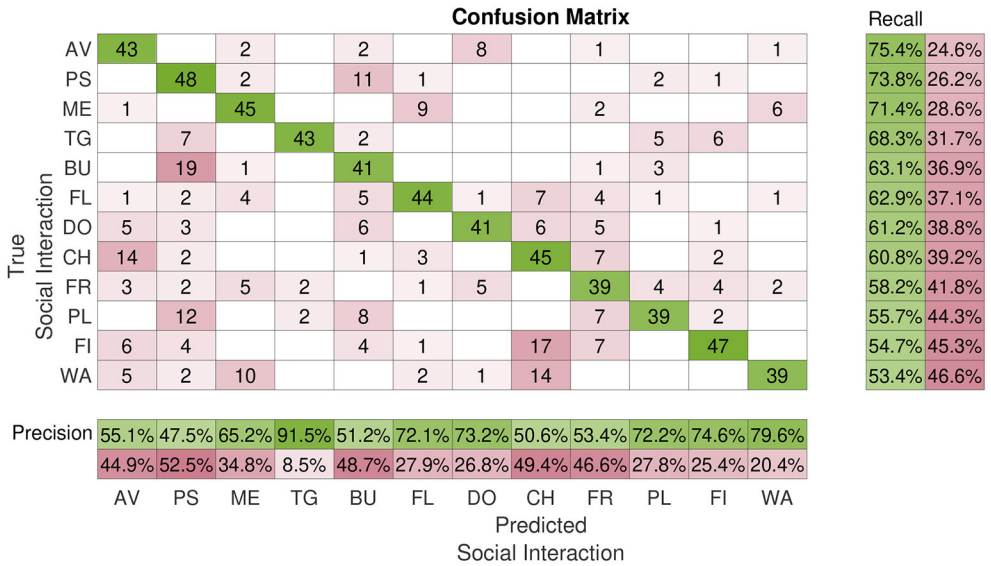

FIGURE 4 | Average classification performance. This figure shows the confusion matrix of the classification experiment (Experiment 2). Rows represent the true interaction class; columns the interaction class reported by the participants in Experiment 2. Matrix entries $m_{i,j}$ report the number of times participants assigned the class j to a video from class i. Rows and columns are sorted by decreasing Recall. AV, avoiding; BU, bumping; CH, chasing; DO, dodging; FI, fighting; FL, flirting; FR, frightening; ME, meeting; PL, pulling; PS, pushing; TG, tug of war; WA, walking.

and Lea, 1994), acceleration (Tremoulet and Feldman, 2000), and speed (Szego and Rutherford, 2008) are the most prominent cues for perceived animacy in psychophysical experiments. Moreover, the blue circle was consistently rated as less animate than the red rectangle [$\beta_1 = -\beta_2 = -8.37\%$, $t_{(299)} = -10$, $p = 1.74 \cdot 10^{-20}$], consistently with the finding that geometrical figures with a body axis are perceived as more animate than those without one, such as circles (Tremoulet and Feldman, 2000).

We further found a significant effect of social interaction on animacy [$F_{(11,299)} = 18.3$, $p = 8.29 \cdot 10^{-28}$]; this suggests that certain classes of social interactions tended to elicit stronger animacy percepts than others. To assess which specific pairs of classes were assigned significantly different animacy rating, we performed *post-hoc F*-tests. This analysis revealed that some classes consistently received higher average animacy ratings: for example, *Fighting* received higher animacy ratings than all other classes [$F_{(1,299)} \geq 24.04$, $p_{adj} \leq 1.03 \cdot 10^{-4}$], with the exception of *Chasing*, which was rated similarly [$F_{(1,299)} = 5.25$, $p_{adj} = 1$]. Analogously, *Bumping* tended to receive lower animacy ratings than all other classes [$F_{(1,299)} \geq 12.44$, $p_{adj} \leq 0.03$], with the exception of *Pushing*, *Frightening*, and *Flirting*, which were rated similarly [$F_{(1,299)} = 8.42$, $p_{adj} \geq 0.26$]. We report in **Figure 3B** all the *post-hoc* F-statistics, and in **Figure 3C** all the corresponding Bonferroni adjusted *p*-values.

3.2. Experiment 2

Figure 4 shows the total confusion matrix M of the classification task. Rows and columns are sorted by decreasing Recall. *Avoiding* was the most accurately classified class by our participants (*Recall* = 75.4%). However, even the hardest class was classified with largely above-chance accuracy (*Walking*: *Recall* = 53.4%; chance level: 8.3%). Nonetheless, there are obviously some misclassifications, especially between *Bumping* and *Pushing* ($m_{BU,PS} = 19$, $m_{PS,BU} = 11$), and between *Fighting* and *Chasing* ($m_{FI,CH} = 17$, $m_{CH,FI} = 2$). These two kinds of mislabeling alone accounted for a large fraction of the total number of mislabelings [$P_{MS}(BU, PS) = 9.8\%$, $P_{MS}(FI, CH) = 6.2\%$].

One possible reason for this misclassification could be the fact that these labels are semantically intrinsically similar and even real videos of these types of social interactions could be mislabeled. This line of reasoning is supported by the fact that in Experiment 1, *Pushing* was the second preferred keyword used to label videos of class *Bumping* (see **Figure 2**). Interestingly, both Precision and Recall (and thus F_1 score) significantly improved across experimental blocks [Precision: $t_{(3)} = 19.5$, $p = 2.93 \cdot 10^{-4}$; Recall: $t_{(3)} = 10.8$, $p = 1.68 \cdot 10^{-3}$; see **Figure 5**]. This indicates a latent learning of the categorization of the classes, which is remarkable since no external feedback about the correctness of the class assignments was provided during the experiment. Such a learning was particularly evident for the following often-confused pairs: *Tug of War* vs. *Pulling*, *Frightening* vs. *Avoiding*, and *Fighting* vs. *Pushing* (not shown).

3.3. Experiment 3

The pairwise semantic distance matrix D is plotted in **Figure 6A**: light shades of green indicate semantically close social interaction classes, while darker shades indicate semantically distant classes. The two pairs associated with the highest mislabeling probability in Experiment 2, *Bumping-Pushing*, and *Fighting-Chasing* [$P_{MS}(BU, PS) = 9.8\%$, $P_{MS}(FI, CH) = 6.2\%$] were generally considered as semantically similar [$d(BU, PU) = 0.49$, $d(FI, CH) = 0.65$]; however, they were not the most similar

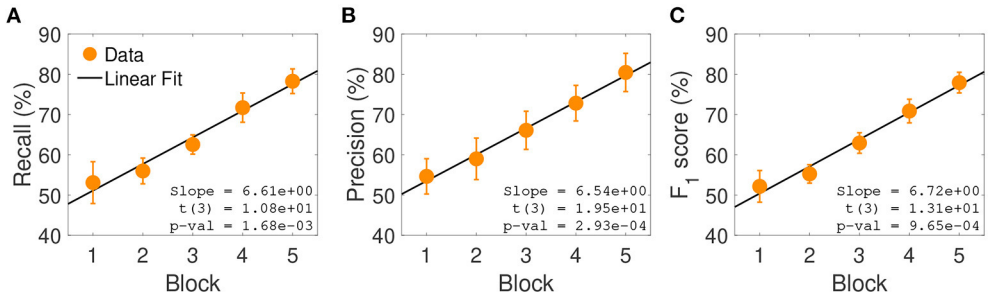

FIGURE 5 | Classification performance across experimental blocks. **(A)** Average block-wise recall. Results are averaged across subjects and social interactions; error bars represent standard errors. Insets show the slope of the estimated linear model, the corresponding t-statistic and p-value. **(B)** Average block-wise precision. **(C)** Average block-wise F_1 score.

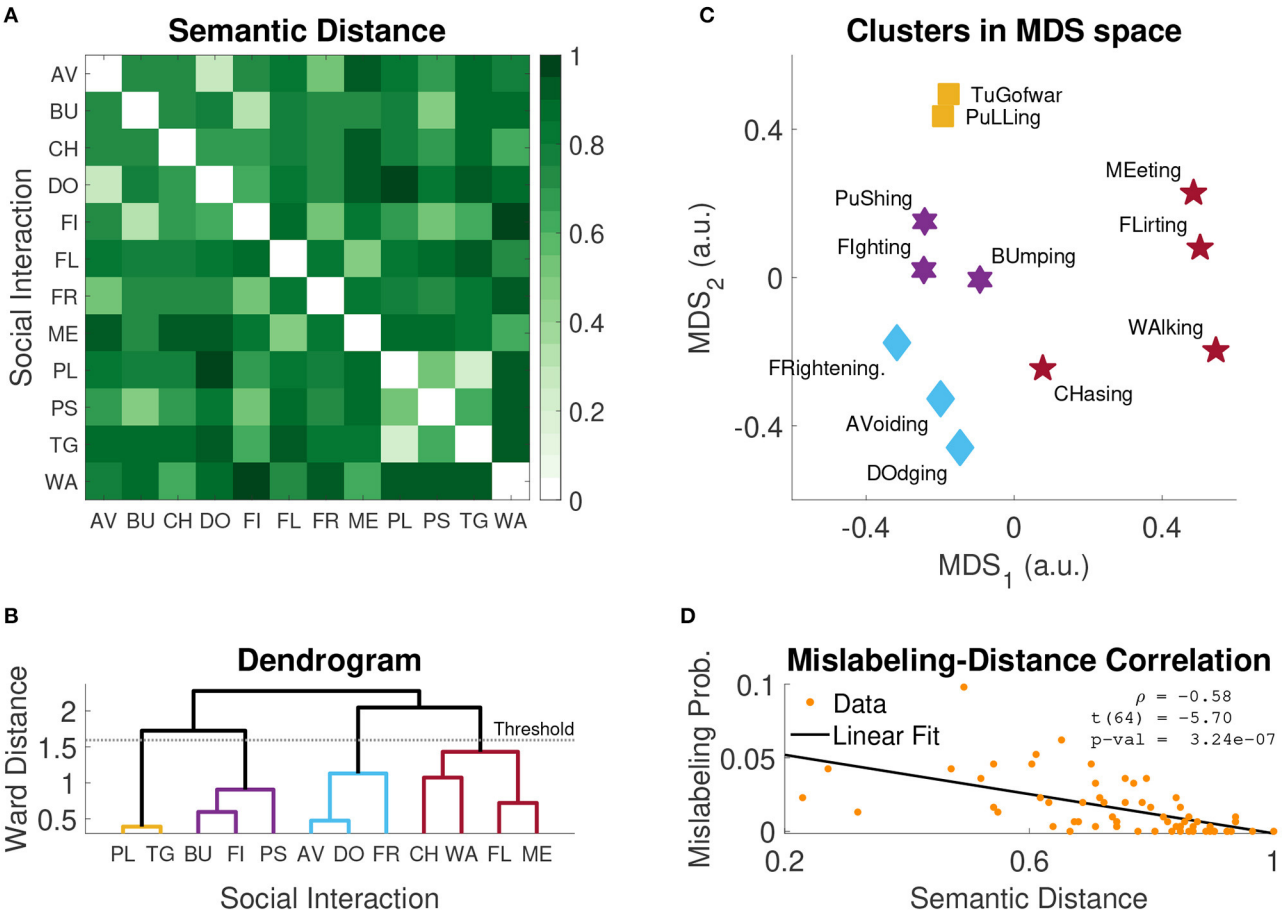

FIGURE 6 | Cluster analysis results. **(A)** Average semantic distances obtained in Experiment 3. **(B)** Dendrogram of hierachical clustering; the horizontal line represents the cut-off threshold used to identify the clusters (i.e., $0.7 * M_{WD}$, where M_{WD} is the maximum Ward distance). **(C)** Clusters of social interactions plotted in low dimensional distance-preserving 2D space identified with Multidimensional Scaling (MDS). **(D)** Average mislabeling probability (Experiment 2) as a function of semantic distance (Experiment 3); inset reports the Pearson's correlation coefficient, the corresponding t-statistic and p-value. AV, avoiding; BU, bumping; CH, chasing; DO, dodging; FI, fighting; FL, flirting; FR, frightening; ME, meeting; PL, pulling; PS, pushing; TG, tug of war; WA, walking.

pairs. Rather, the three most semantically similar pairs were *Pulling-Tug of War*, *Avoiding-Dodging*, and *Bumping-Fighting* [$d(PU, TG) = 0.23$, $d(AV, DO) = 0.27$, $d(BU, FI) = 0.32$]. Nevertheless, regardless of this apparent discrepancy for these few extreme examples, mislabeling probability $P_{MS}(i, j)$ and semantic distance $d(i, j)$ were significantly anti-correlated [$\rho =$

-0.58, $t_{(64)} = -5.7$, $p = 3.24 \cdot 10^{-7}$; **Figure 6D**]; this suggests that the more semantically similar two social interaction classes are, the more likely they are of being confused in a video labeling task.

Multidimensional scaling (MDS) provides a compact 2D visualization of the semantic similarity space (**Figure 6C**). Since MDS is inherently spatial, items that were rated as being highly similar are spatially close to each other in the final map. The map effectively shows which classes of social interactions are semantically similar and which are not. For example, let us consider the hypothetical groups G_1 ={Tug of War, Pulling} and G_2 ={Frightening, Avoiding, Dodging}. Participants recognized that *Tug of War* and *Pulling* involve similar interactions between the agents, and that these interactions are different from those occurring in the classes *Frightening*, *Avoiding*, and *Dodging*. For this reason, participants tended to assign high pairwise similarity scores to intra-group pairs, and low to inter-group pairs. This pattern of scoring is captured by MDS and evident in the resulting map (**Figure 6C**).

The agglomerative hierarchical cluster analysis on the distance matrix D (**Figure 6B**) confirms this intuition and identifies four distinct semantic clusters; such clusters are visualized in the MDS map with four different symbols (**Figure 6C**). This analysis supports the conclusion that misclassified labels tend to belong to the same semantic cluster. While not all misclassifications can be explained by semantic similarity, many confusions can be accounted for by this factor. For example, *Pushing* vs. *Bumping*, *Walking* vs. *Meeting*, *Avoiding* vs. *Dodging*.

To summarize, our analysis of semantic similarity shows that many of the labeling confusions observed in Experiment 2 can be explained by the semantic similarity of the class labels.

4. DISCUSSION

In this work, we introduced a novel framework for the automatic generation of videos of socially interacting virtual agents. The underlying model is a nonlinear dynamical system that specifies heading direction and forward speed of the agents. Our model is able to generate as many as 15 different interaction classes, defined by different parameter sets. We validated our model with three different behavioral experiments, in which participants were able to consistently identify the intended interaction classes. Our model is thus suitable for the automatic generation of animations of socially interacting agents. Furthermore, the generation process is also amenable to full parametric control. This feature allows the creation of highly-controlled and arbitrarily-large datasets for in-depth psychophysical and electrophysiological characterization of the perception of social interactions. The model thus overcomes the major limitations that come with hand-crafted, hard-coded, rule-based, and real-video-based approaches (1) to visual stimuli generation. Importantly, the generative nature of the model, makes it a valuable tool also for the development of mechanistic and neural *decoder* models of social perception: model responses to the heterogeneous set of highly-controlled social stimuli here introduced can be rigorously tested for the development of more accurate and brain-like decoder models that replicate human behavioral and neural responses. Recent work (Shu et al., 2018, 2019, 2020), aimed at building a mechanistic model of social inference, used a similar approach.

Shu et al. (2019, 2020) also proposed generative models of social interactions. Unlike the ones proposed in these studies, the generative model introduced in this work does not directly lend itself to the study of the interactions between intuitive physics and social inferences (Battaglia et al., 2013). However, substantial evidence suggests that physical and social judgments are mediated by different brain regions (Isik et al., 2017; Sliwa and Freiwald, 2017). More importantly, our model is not limited to describing cooperative and obstructive behaviors and thus seems better suited to study more general social interaction classes.

The identification of suitable parameters for the classes modeled in this work was not automatic: it was conducted using a simulation-based heuristic procedure. This is an obvious limitation of our work. Nevertheless, once the parameters are available, they can be used to automatically generate arbitrary numbers of coupled trajectories for each interaction class (by randomly sampling initial conditions, via-points, and noise). With this procedure, we were able to find suitable parameters for only 15 specific interaction classes. However, to the best of our knowledge, no other method is able to automatically generate more than a handful of individual or socially-interactive behaviors (Blackwell, 1997; Paris et al., 2007; Luo et al., 2008; Russell et al., 2017; Shu et al., 2019, 2020). Future work will extend the range of modeled classes by using system identification methods (e.g., Schön et al., 2011; Gao et al., 2018; Gonçalves et al., 2020) to automatically extract model parameters from preexisting trajectories—extracted, for example, from real videos.

Another possible limitation of our work is that all our participants were recruited from a German university; while this might, in theory, represent a biased sample, previous studies (Rimé et al., 1985) suggest that the perception of social interactions from impoverished stimuli is a phenomenon that is highly stable across cultures. Specifically, these authors showed that African, European, and Northern American participants provided similar interpretations to animated videos of geometrical shapes. This suggests that our findings would not have significantly changed if we had recruited a more heterogeneous sample.

In this work, we used the trajectories generated by our model to animate simple geometrical figures. The resulting abstract visual stimuli can be directly applied to characterize the kinematic features underlying the inference of social interactions. However, the trajectories can also be used as a basis for richer visual stimuli. For example, in ongoing work, we have been developing methods to link the speed and direction dynamics generated by the model to articulating movements of three-dimensional animal models. This approach allows the generation of highly controlled and realistic videos of interacting animals, which can be used to study social interaction perception in the corresponding animal models with ecologically valid stimuli. Furthermore,

contrasting the neural responses to impoverished and realistic visual stimuli can help identify the brain regions and neural computations mediating the extraction of the relevant kinematic features and the subsequent construction of social percepts.

Finally, even though the proposed model is mainly aimed to provide a tool to facilitate the design of in-depth psychophysical and electrophysiological studies of social interaction perception, we speculate that it can also be helpful in the development of machine vision systems for the automatic detection of social interactions. Specifically, the development of effective modern machine vision systems tends to be heavily dependent on the availability of large numbers of appropriately-labeled videos of social interactions (Rodríguez-Moreno et al., 2019; Stergiou and Poppe, 2019). A popular approach to this problem is to use clips extracted from already existing (YouTube) videos and movies. However, one of the reasons why feature-based (e.g. Kumar and John, 2016; Sehgal, 2018) and especially deep-neural-network-based (e.g., Karpathy et al., 2014; Carreira and Zisserman, 2017; Gupta et al., 2018) vision systems require *big data* is that they need to learn to ignore irrelevant information that is inevitably present in real videos. Therefore, we hypothesize that pre-training such systems with stylized videos of socially interacting agents—such as the very same generated by our model or appropriate avatar-based extensions—might greatly reduce their training time and possibly improve their performance. Future work will test this hypothesis.

To sum up, this work introduced a novel generative model of social interactions. The results of our psychophysical experiments suggest that the model is suitable for the automatic generation of arbitrarily-numerous and highly-controlled videos of socially interacting agents for comprehensive studies of animacy and social interaction perception. Our model can also be potentially used to create large, noise-free, and annotated datasets that can facilitate the development of mechanistic and neural models of social perception, as well as the design of machine vision systems for the automatic recognition of human interactions.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics board of the University of Tübingen. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the German Federal Ministry of Education and Research (BMBF FKZ 01GQ1704), the Human Frontiers Science Program (HFSP RGP0036/2016), the German Research Foundation (DFG GZ: KA 1258/15-1), and the European Research Council (ERC 2019-SyG-RELEVANCE-856495).

ACKNOWLEDGMENTS

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting AS. The authors would also like to thank the participants who took part in the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2021.648527/full#supplementary-material

REFERENCES

Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition* 113, 329–349. doi: 10.1016/j.cognition.2009.07.005

Barrett, H. C., Todd, P. M., Miller, G. F., and Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: a cross-cultural study. *Evol. Hum. Behav*. 26, 313–331. doi: 10.1016/j.evolhumbehav.2004.08.015

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci. U.S.A*. 110, 18327–18332. doi: 10.1073/pnas.1306572110

Bicho, E., Mallet, P., and Schöner, G. (2000). Target representation on an autonomous vehicle with low-level sensors. *Int. J. Robot. Res*. 19, 424–447. doi: 10.1177/02783640022066950

Bicho, E., and Schöner, G. (1997). The dynamic approach to autonomous robotics demonstrated on a low-level vehicle platform. *Robot. Auton. Syst*. 21, 23–35. doi: 10.1016/S0921-8890(97)00004-3

Blackwell, P. (1997). Random diffusion models for animal movement. *Ecol. Model*. 100, 87–102. doi: 10.1016/S0304-3800(97)00153-1

Blythe, P. W., Todd, P. M., and Miller, G. F. (1999). "How motion reveals intention: categorizing social interactions," in *Simple Heuristics That Make Us Smart*, eds G. Gigerenzer and P. M. Todd (Oxford University Press). pp. 257–285.

Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 6299–6308. doi: 10.1109/CVPR.2017.502

Castelli, F., Frith, C., Happé, F., and Frith, U. (2002). Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* 125, 1839–1849. doi: 10.1093/brain/awf189

Castelli, F., Happé, F., Frith, U., and Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* 12, 314–325. doi: 10.1006/nimg.2000.0612

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition* 107, 705–717. doi: 10.1016/j.cognition.2007.08.001

Dittrich, W. H., and Lea, S. E. (1994). Visual perception of intentional motion. *Perception* 23, 253–268. doi: 10.1068/p230253

Fajen, B. R., and Warren, W. H. (2003). Behavioral dynamics of steering, obstacle avoidance, and route selection. *J. Exp. Psychol.* 29:343. doi: 10.1037/0096-1523.29.2.343

Freiwald, W. A. (2020). The neural mechanisms of face processing: cells, areas, networks, and models. *Curr. Opin. Neurobiol.* 60, 184–191. doi: 10.1016/j.conb.2019.12.007

Gao, S., Zhou, M., Wang, Y., Cheng, J., Yachi, H., and Wang, J. (2018). Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 601–614. doi: 10.1109/TNNLS.2018.2846646

Gao, T., McCarthy, G., and Scholl, B. J. (2010). The wolfpack effect: perception of animacy irresistibly influences interactive behavior. *Psychol. Sci.* 21, 1845–1853. doi: 10.1177/0956797610388814

Gao, T., Newman, G. E., and Scholl, B. J. (2009). The psychophysics of chasing: a case study in the perception of animacy. *Cogn. Psychol.* 59, 154–179. doi: 10.1016/j.cogpsych.2009.03.001

Gelman, R., Durgin, F., and Kaufman, L. (1995). "Distinguishing between animates and inanimates: not by motion alone," in *Causal Cognition: A Multidisciplinary Debate*, eds, D, Sperber, D. Premack, and A. J. Premack (Oxford: Clarendon Press), 150–184. doi: 10.1093/acprof:oso/9780198524021.003.0006

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., et al. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife* 9:e56261. doi: 10.7554/eLife.56261

Gordon, A. S., and Roemmele, M. (2014). "An authoring tool for movies in the style of Heider and Simmel," in *International Conference on Interactive Digital Storytelling* (Singapore: Springer), 49–60. doi: 10.1007/978-3-319-12337-0_5

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., and Alahi, A. (2018). "Social GAN: socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 2255–2264. doi: 10.1109/CVPR.2018.00240

Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. *Am. J. Psychol.* 57, 243–259. doi: 10.2307/1416950

Isik, L., Koldewyn, K., Beeler, D., and Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proc. Natl. Acad. Sci. U.S.A.* 114, E9145–E9152. doi: 10.1073/pnas.1714471114

Kaduk, K., Elsner, B., and Reid, V. M. (2013). Discrimination of animate and inanimate motion in 9-month-old infants: an ERP study. *Dev. Cogn. Neurosci.* 6, 14–22. doi: 10.1016/j.dcn.2013.05.003

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (Columbus, OH), 1725–1732. doi: 10.1109/CVPR.2014.223

Kerr, W., and Cohen, P. (2010). "Recognizing behaviors and the internal state of the participants," in *2010 IEEE 9th International Conference on Development and Learning* (Ann Arbor, MI), 33–38. doi: 10.1109/DEVLRN.2010.5578868

Kumar, S. S., and John, M. (2016). "Human activity recognition using optical flow based feature set," in *2016 IEEE International Carnahan Conference on Security Technology (ICCST)* (Orlando, FL), 1–5. doi: 10.1109/CCST.2016.7815694

Luo, L., Zhou, S., Cai, W., Low, M. Y. H., Tian, F., Wang, Y., et al. (2008). Agent-based human behavior modeling for crowd simulation. *Comput. Anim. Virt. Worlds* 19, 271–281. doi: 10.1002/cav.238

McAleer, P., Kay, J. W., Pollick, F. E., and Rutherford, M. (2011). Intention perception in high functioning people with autism spectrum disorders using animacy displays derived from human actions. *J. Autism Dev. Disord.* 41, 1053–1063. doi: 10.1007/s10803-010-1130-8

McAleer, P., and Pollick, F. E. (2008). Understanding intention from minimal displays of human activity. *Behav. Res. Methods* 40, 830–839. doi: 10.3758/BRM.40.3.830

Michotte, A. (1946). *The Perception of Causality, Vol. 21*. New York, NY: Basic Books.

Mukovskiy, A., Slotine, J.-J. E., and Giese, M. A. (2013). Dynamically stable control of articulated crowds. *J. Comput. Sci.* 4, 304–310. doi: 10.1016/j.jocs.2012.08.019

Oatley, K., and Yuill, N. (1985). Perception of personal and interpersonal action in a cartoon film. *Br. J. Soc. Psychol.* 24, 115–124. doi: 10.1111/j.2044-8309.1985.tb00670.x

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C.-C., et al. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition* 130, 360–379. doi: 10.1016/j.cognition.2013.11.011

Paris, S., Pettré, J., and Donikian, S. (2007). "Pedestrian reactive navigation for crowd simulation: a predictive approach," in *Computer Graphics Forum, Vol. 26* (Prague: Wiley Online Library), 665–674. doi: 10.1111/j.1467-8659.2007.01090.x

Reichardt, W., and Poggio, T. (1976). Visual control of orientation behaviour in the fly: Part I. A quantitative analysis. *Q. Rev. Biophys.* 9, 311–375. doi: 10.1017/S0033583500002523

Reimann, H., Iossifidis, I., and Schöner, G. (2011). "Autonomous movement generation for manipulators with multiple simultaneous constraints using the attractor dynamics approach," in *2011 IEEE International Conference on Robotics and Automation* (Shanghai), 5470–5477. doi: 10.1109/ICRA.2011.5980184

Rimé, B., Boulanger, B., Laubin, P., Richir, M., and Stroobants, K. (1985). The perception of interpersonal emotions originated by patterns of movement. *Motiv. Emot.* 9, 241–260. doi: 10.1007/BF00991830

Rodríguez-Moreno, I., Martínez-Otzeta, J. M., Sierra, B., Rodriguez, I., and Jauregi, E. (2019). Video activity recognition: state-of-the-art. *Sensors* 19:3160. doi: 10.3390/s19143160

Roether, C. L., Omlor, L., Christensen, A., and Giese, M. A. (2009). Critical features for the perception of emotion from gait. *J. Vis.* 9:15. doi: 10.1167/9.6.15

Russell, J. C., Hanks, E. M., Modlmeier, A. P., and Hughes, D. P. (2017). Modeling collective animal movement through interactions in behavioral states. *J. Agric. Biol. Environ. Stat.* 22, 313–334. doi: 10.1007/s13253-017-0296-3

Saxe, R., and Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind." *Neuroimage* 19, 1835–1842. doi: 10.1016/S1053-8119(03)00230-1

Scholl, B. J., and Gao, T. (2013). "Perceiving animacy and intentionality: visual processing or higher-level judgment," in *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention* eds, M. D. Rutherford and V. A. Kuhlmeier (Cambridge, MA: MIT Press), 197–230. doi: 10.7551/mitpress/9780262019279.003.0009

Scholl, B. J., and Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends Cogn. Sci.* 4, 299–309. doi: 10.1016/S1364-6613(00)01506-0

Schön, T. B., Wills, A., and Ninness, B. (2011). System identification of nonlinear state-space models. *Automatica* 47, 39–49. doi: 10.1016/j.automatica.2010.10.013

Schöner, G., and Dose, M. (1992). A dynamical systems approach to task-level system integration used to plan and control autonomous vehicle motion. *Robot. Auton. Syst.* 10, 253–267. doi: 10.1016/0921-8890(92)90004-I

Schöner, G., Dose, M., and Engels, C. (1995). Dynamics of behavior: theory and applications for autonomous robot architectures. *Robot. Auton. Syst.* 16, 213–245. doi: 10.1016/0921-8890(95)00049-6

Schultz, J., and Bülthoff, H. H. (2019). Perceiving animacy purely from visual motion cues involves intraparietal sulcus. *NeuroImage* 197, 120–132. doi: 10.1016/j.neuroimage.2019.04.058

Sehgal, S. (2018). "Human activity recognition using BPNN classifier on hog features," in *2018 International Conference on Intelligent Circuits and Systems (ICICS)* (Phagwara), 286–289. doi: 10.1109/ICICS.2018.00065

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* 27, 125–140. doi: 10.1007/BF02289630

Shepard, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika* 27, 219–246. doi: 10.1007/BF02289621

Shu, T., Kryven, M., Ullman, T. D., and Tenenbaum, J. B. (2020). "Adventures in flatland: perceiving social interactions under physical dynamics," in *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (Toronto).

Shu, T., Peng, Y., Fan, L., Lu, H., and Zhu, S.-C. (2018). Perception of human interaction based on motion trajectories: from aerial videos to decontextualized animations. *Top. Cogn. Sci.* 10, 225–241. doi: 10.1111/tops.12313

Shu, T., Peng, Y., Lu, H., and Zhu, S. (2019). "Partitioning the perception of physical and social events within a unified psychological space," in *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (Montreal).

Sliwa, J., and Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. *Science* 356, 745–749. doi: 10.1126/science.aam6383

Springer, K., Meier, J. A., and Berry, D. S. (1996). Nonverbal bases of social perception: developmental change in sensitivity to patterns of motion that reveal interpersonal events. *J. Nonverb. Behav*. 20, 199–211. doi: 10.1007/BF02248673

Stergiou, A., and Poppe, R. (2019). Analyzing human-human interactions: a survey. *Comput. Vis. Image Understand*. 188:102799. doi: 10.1016/j.cviu.2019.102799

Szego, P. A., and Rutherford, M. D. (2008). Dissociating the perception of speed and the perception of animacy: a functional approach. *Evol. Hum. Behav*. 29, 335–342. doi: 10.1016/j.evolhumbehav.2008.04.002

Thurman, S. M., and Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PLoS ONE* 9:e112539. doi: 10.1371/journal.pone.0112539

Träuble, B., Pauen, S., and Poulin-Dubois, D. (2014). Speed and direction changes induce the perception of animacy in 7-month-old infants. *Front. Psychol*. 5:1141. doi: 10.3389/fpsyg.2014.01141

Tremoulet, P. D., and Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception* 29, 943–951. doi: 10.1068/p3101

Tremoulet, P. D., and Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Percept. Psychophys*. 68, 1047–1058. doi: 10.3758/BF03193364

Troje, N., Simion, F., Bardi, L., Mascalzoni, E., Regolin, L., Grossman, E., et al. (2013). *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention*. Cambridge, MA: MIT Press.

van Buren, B., Gao, T., and Scholl, B. J. (2017). What are the underlying units of perceived animacy? Chasing detection is intrinsically object-based. *Psychon. Bull. Rev*. 24, 1604–1610. doi: 10.3758/s13423-017-1229-4

van Buren, B., Uddenberg, S., and Scholl, B. J. (2016). The automaticity of perceiving animacy: goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychon. Bull. Rev*. 23, 797–802. doi: 10.3758/s13423-015-0966-5

Walbrin, J., Downing, P., and Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia* 112, 31–39. doi: 10.1016/j.neuropsychologia.2018.02.023

Ward, J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc*. 58, 236–244. doi: 10.1080/01621459.1963.10500845

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A*. 111, 8619–8624. doi: 10.1073/pnas.1403112111

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.