Better Hooks to Catch Better Fish:
Using Artificial Intelligence to Study Group Dynamics at Massive Scale

Jesse Hoey
University of Waterloo, Canada

Tobias Schröder and Jonathan Morgan
Potsdam University of Applied Sciences, Germany

Kimberly B. Rogers
Dartmouth College, United States

Deepak Rishi and Meiyappan Nagappan
University of Waterloo, Canada

Author Note

Hoey and Schröder share first authorship. We thank Hong-Mao Li for help with data analysis. We acknowledge support from the Trans-Atlantic Platform's Digging into Data Program (funding provided by NSERC and SSHRC in Canada, DFG in Germany, and NSF in the United States).

Address correspondence to Jesse Hoey, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. E-mail: jesse.hoey@uwaterloo.ca

Abstract

There are two central problems in studying group dynamics. First, because empirical research on groups relies on manual coding, it is hard to study groups in large numbers (the scaling problem). Second, conventional methods in behavioural science are based on the general linear model, which fails to capture the often nonlinear interaction dynamics occurring in small groups (the dynamics problem). We discuss technological advances in artificial intelligence that might help overcome these limitations. Machine learning helps to address the scaling problem, as massive computing power can be harnessed to multiply manual codings of group interactions. Computer simulations help to address the dynamics problem by implementing social psychological theory in data-generating algorithms that allow for sophisticated statements and tests of such theory. As an illustration of these developments in small group research, we describe an ongoing research project aimed at computational analysis of virtual software development teams.

Keywords: Artificial Intelligence, Group Dynamics, Machine Learning, Social Simulation, Virtual Teams

Better Hooks to Catch Better Fish:
Using Artificial Intelligence to Study Group Dynamics at Massive Scale

The digital transformation of society has opened up novel opportunities for studying social interaction. People increasingly use digital tools and social-network platforms to communicate with each other, producing large amounts of digital data that can be analyzed to answer psychological or sociological questions. Often, such data can be secured by researchers through data mining; executing a few lines of code on a personal computer provides an alluring alternative to the cumbersome process of creating scientifically-relevant data through surveys, observations, or experiments. Yet, as progress often comes with a price tag, the analysis of such readily available data can be very challenging for at least two reasons. One is the sheer volume of data, which requires complex database technologies and increased computational resources in comparison to classic quantitative methods in social research. Another challenge is that social media data were not created with the purpose of answering the social scientist's specific research questions. The analysis of these data is ultimately a qualitative exercise, requiring skillful interpretation of naturally-occurring social interactions. However, the amount of data makes traditional observation, categorization, and interpretation methods impossible to carry out at the scale required. Fortunately, researchers in artificial intelligence (AI) have developed powerful algorithms that can be leveraged to help social scientists tackle the challenges involved in the analysis of "big" behavioural data.

These developments have been described with much excitement under the label "computational social science" (Lazer et al., 2009). While there is no lack of big claims linking computational social science to groundbreaking inventions such as the telescope in physics or the microscope in biology and, consequently, predicting traditional research methods to become obsolete (e.g., Savage & Burrows, 2007), we believe the hopes and promises of these methods are currently much bigger than the actual achievements (cf. Couper, 2013; Schober, Pasek, Guggenheim, Lampe, & Conrad, 2016). One of the main

problems, in our view, is a substantial disconnect between disciplines such as computer science or physics, which produce the methodological expertise for large-scale data analysis and modeling complex social systems, and disciplines such as psychology or sociology, which provide a rich landscape of theories and empirical evidence that enable thoughtful use of these methods. Developing a meaningful "computational social psychology" (Vallacher, Read, & Nowak, 2017) that capitalizes on novel technologies to advance the existing body of knowledge about social interaction will require diligent work and cross-disciplinary interaction for years to come.

The purpose of the present paper is to discuss the promises of a cross-disciplinary, computational approach to the study of small-group dynamics, and describe how such an approach might proceed using our own research as an example. Importantly, we not only want to review computational methods for using large amounts of social media data, but also point out the necessity and feasibility of doing so in a theoretically-informed way. To use a metaphor, we want to dig into digital group-dynamics data with a sophisticated, artificially intelligent shovel that "knows" about social psychology. To show how this is possible, we will briefly review our own work in developing Bayesian Affect Control Theory, which mathematically integrates widely-accepted psychological and sociological theories of social interaction and thus enables us to create artificially intelligent agents that are aware of social scientific knowledge (Hoey, Schröder, & Alhothali, 2016; Schröder, Hoey, & Rogers, 2016). We also show some preliminary results from an ongoing research project to illustrate the logic and feasibility of small-group research enhanced by artificial intelligence (AI).

Before we turn to our own work in this field, however, we provide a brief review of developments in AI and computational social science that are relevant, in our opinion, for group-dynamics research in general. We structure this review around two persistent problems in the field that we believe can be addressed in fundamentally novel ways with AI methods. The first we call the scaling problem: it has been a cumbersome task in the past

to gather and interpret data on small-group dynamics; hence, limited resources often prevent the execution of this kind of research at massive scale. The second is the dynamics problem: small groups are complex systems whose analysis requires more sophisticated mathematical tools than the general linear model widely taught in the social and behavioural sciences; hence, we often lack an understanding of the deep information-processing mechanisms at the heart of group dynamics. Both challenges can be met with the novel tools under development in research on artificial intelligence.

The Scaling Problem

In comparison to social psychological studies focusing on individuals, small-groups research is difficult to scale. Owing to the statistical non-independence of data from individual members of one group, the permissible level of analysis for many questions is the group, not the individual, resulting in a need for much larger sample sizes than the typical study of individuals would require (cf. Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). In addition, coordination problems related to physical co-presence abound – one cannot simply study 1,000 groups in an online survey. Finally, if one is interested in the micro-dynamics of a group, a very fine time-resolution of the ongoing interaction is often required, further increasing the economic demands of generating high-quality datasets. While these issues have traditionally been resolved through manual coding of behaviour within a small number of closely-monitored groups, we believe that existing AI technology can be harnessed to overcome the scaling problem.

Coding Schemes for Observation of Small Groups

While the entire methodological apparatus of the social and behavioural sciences has been used to study group dynamics (for recent reviews, see Forsyth, 2018; Kerr & Tindale, 2014), one of the most distinguished approaches (and perhaps the approach most seriously plagued by the scaling problem) has arguably been the systematic observation of group interactions. Scholars studying group dynamics via direct observation have developed

elaborate systems of categories, purporting to represent all the possible types of communicative acts that one group member can direct to another. These observation systems come with specific instructions and manuals for how to divide the ongoing flow of interaction into discrete segments and assign to each of these segments one or several categories (e.g. Bales, 1950; Schermuly, Schröder, Nachtwei, & Scholl, 2010). While systematic observation at the level of the act provides an invaluable data source for studying the inner mechanics of groups, the task is daunting. Estimations of the time required to apply these coding schemes range from 8 up to 50 hours per coder for each hour of group interaction, depending on the amount of detail taken into account (Schermuly & Schölmerich, 2017). As a consequence, researchers only study small numbers of groups, small samples of the total interactions, or shy away from studying groups at such a level of detail altogether (cf. Kerr & Tindale, 2014). It would be impossible with this method to study, say, 1,000 teams of software developers and analyse each of their communicative acts over the course of an entire year.

As a specific – and possibly the pioneering – example of categorization systems for group behaviour, consider Interaction Process Analysis (IPA), developed by Bales (1950). As displayed in Table 1, acts directed from one member of a group to another can belong to one of twelve different functional categories. These categories cluster together into two types of task-oriented behaviours (giving vs. soliciting information or guidance) and two types of expressive behaviours (positive vs. negative) aimed at socio-emotional regulation. IPA and similar categorical systems have been employed in numerous studies reviewed by Bales (1999) to pursue questions such as status emergence in groups, over-time phases in group dynamics, and the effectiveness of collective problem-solving.

Automatic Categorization with Machine Learning

Formally, the task required of human coders using a system like IPA is to map a complex sensory input (consisting of soundwaves corresponding to their verbal and vocal

Table 1
Categories of Interaction Process Analysis (IPA) (Bales, 1999, p. 165).

Problem Areas	Observation Categories
Expressive, Social-Emotional, Positive Reactions	1. *Shows solidarity*, raises other's status, gives help, reward 2. *Shows tension release*, jokes, laughs, shows satisfaction 3. *Agrees*, shows passive acceptance, understands, concurs, complies
Instrumental-Adaptive, Task Orientation, Attempted Answers	4. *Gives suggestion*, direction, implying autonomy for other 5. *Gives opinion*, evaluation, analysis, expresses feeling, wish 6. *Gives orientation*, information, repeats, clarifies, confirms
Instrumental-Adaptive, Task Orientation, Questions	7. *Asks for orientation*, information, repetition, confirmation 8. *Asks for opinion*, evaluation, analysis, expression of feeling 9. *Asks for suggestion*, direction, possible ways of action
Expressive, Social-emotional, Negative Reactions	10. *Disagrees*, shows passive rejection, formality, withholds help 11. *Shows tension*, asks for help, withdraws out of field 12. *Shows antagonism*, deflates other's status, defends or asserts self

acts, and of photons carrying information about facial and gestural expressions) to a narrow set of well-defined analytical categories. With recent advances in machine learning (for review, see Jordan & Mitchell, 2015), artificially intelligent agents have been built that are as good as humans at this kind of task or even outperform them, including in psychological rating tasks such as inferring sexual orientation from facial features (Wang & Kosinski, 2018). A key technology that has fuelled recent success in this area of AI is called "Deep Learning" (Goodfellow, Bengio, & Courville, 2016; LeCun, Bengio, & Hinton, 2015).

As visualized in Figure 1, Deep Learning is implemented in an artificial neural network with many layers. Abstracting from learning principles of the human brain, such a network "learns" by changing the connection weights and thus the flow of activation between individual neurons. What is "learned" is a mapping from a high dimensional input (such as images) to a low-dimensional output (such as categories). The exact mechanisms, or algorithms, by which the connection weights change are the subject of much research in AI, but a common feature of such algorithms is that the networks need to be trained. This usually happens by showing the network a sample of mappings between physical inputs

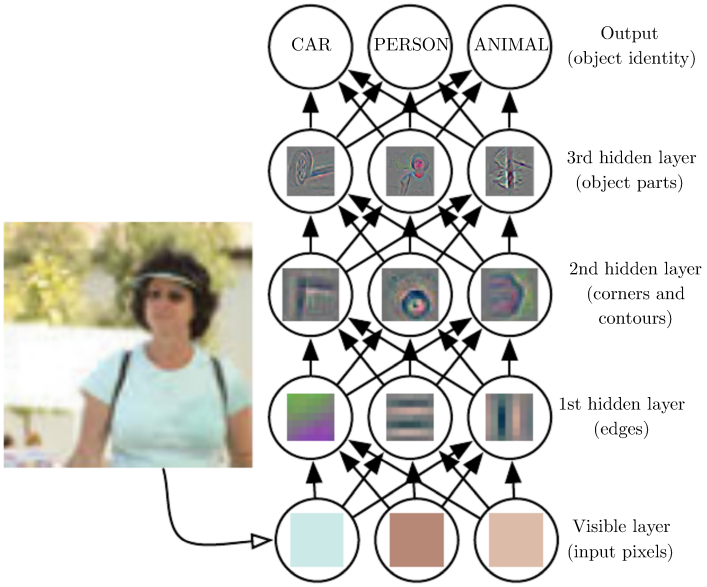

Figure 1. Simple example for categorizing physical features into symbolic concepts with a deep learning network. Each circle represents an aritifical neuron, and each arrow carries a "weight" corresponding to the synaptic strength of the connection between neurons. Only a tiny subset of the total number of neurons are shown, as a real deep network usually has tens of thousands of them at each layer.

and categorical outputs. These samples often need to be provided by human annotators, an issue revisited later in this section. Once a network has learned the relevant mapping, it is able to extract high-level features and categories from an input pattern and generalize the mapping to inputs it has not seen before. In Figure 1, the displayed network solves the problem of image recognition – i.e., identify a categorical object from a large vector of pixels.

While visual object recognition is probably one of the most successful applications of deep learning to date (e.g., this forms the basis of image search online and is an important problem to be solved by self-driving cars), the technology can be used in principle to map any complex set of features to a system of categories. For example, it is conceivable (although we are not aware of any existing implementation of this idea) to have a deep-learning network infer Bales' IPA categories (see Table 1) from a video recording of a group discussion. Human coders knowledgable about IPA would still be required to train

the network, but once that job is completed, the resulting learned deep network could be used as an automated assistant that rates group interactions as they occur, potentially providing insights to the group and helping steer it towards a more cooperative dynamic.

Natural Language Processing and Sentiment Analysis

Another area of recently flourishing AI research, partly overlapping with machine learning, is Natural Language Processing (NLP), the automated analysis of text written by humans (for review, see Hirschberg & Manning, 2015). This technology is relevant for small-group research because group collaboration has become increasingly virtual, at least in parts, through commercial project management software and social media platforms such as GitHub (github.com), where millions of people collaborate online to develop software and other artifacts. Virtual collaboration means that group members interact by sending each other text messages, which are stored on the platform and often accessible to researchers. Similarly to video or images, written data could be be used to infer functional group-interaction categories. In this case, the input provided to a deep network consists of raw text found on social media platforms, while the output is again a set of categorical labels such as IPA categories.

A relevant application of NLP is sentiment analysis, the mapping of linguistic written text to the evaluative sentiment expressed in that text (e.g. “good” vs. “bad”, or “like” vs. “dislike”) (e.g., Medhat, Hassan, & Korashy, 2014; Pang & Lee, 2008). Sentiment analysis is potentially useful to the study of affective dynamics in virtual group collaboration beyond a functional classification of acts. Algorithms can be built, for example, to identify specific words that carry direct evaluative meanings (e.g., *hate* or *love*), evoke implicit affective meanings (Osgood, 1962) (e.g., most people associate positive feelings with a *child* and negative feelings with a *rapist*), or make use of subtle linguistic signs (e.g., exclamation marks or emoji). Sentiment analysis of social media data is a dynamic and expanding research field, which has studied everything from product reviews through happiness

research and election forecasting (e.g., Alhothali & Hoey, 2015; L. Mitchell, Frank, Harris, Dodds, & Danforth, 2013; Pang & Lee, 2008; Pletea, Vasilescu, & Serebrenik, 2014a; Tumasjan, Sprenger, Sandner, & Welpe, 2010). However, to our knowledge, most if not all existing social psychological applications of sentiment analysis are largely atheoretical data-mining exercises. In contrast, we believe that current theory and model-driven artificial intelligence enable us to go beyond digital fishing expeditions and develop theory-driven research agendas that increase our understanding of the mechanisms underlying dynamics in small groups. Rather than using a "more hooks to catch more fish" approach, we are using sophisticated technology grounded in social-psychological theory to more precisely locate the prize catches.

Overcoming the Scaling Problem

Current data-driven machine learning methods have not solved the scaling problem entirely: labeled training data is still required at a massive scale, and this requires extensive annotation work by humans. The reason for this is that neural network models involve a massive number of parameters (one "weight" for each "neuron") that must be tuned or learned from data. As in any standard regression problem, the number of data points needed must be on the same order as the number of parameters in order to guarantee successful learning. Two methods can be used to overcome this challenge. First, it is possible to use existing social media data as labeled examples so long as one can identify signals that correspond to the labels which are explicitly included in the data itself. For example, emoji can be used as direct labels of sentiment (Felbol, Mislove, Søgaard, Rahwan, & Lehmann, 2017). Unsupervised methods can also be used to characterise dimensions of meaning by learning statistical patterns in large document corpora (Blei, 2012; Kozlowski, Taddy, & Evans, 2018). Further, data labeling processes can be "gamified" to make them a part of the task naturally being worked on by human social media users as they interact online (Law, Gajos, Wiggins, Gray, & Williams, 2017). These

so-called “crowdsourcing” methods allow labels to be obtained as an organic byproduct of human work processes, using only simple adjustments to interfaces or mechanisms.

The second method is to use “top-down” approaches to limit the scope of the models being learned. Rather than searching over the space of all possible settings of deep network parameters, for example, one can focus on the specific settings that are likely to yield good results when trained with small amounts of data. This can be accomplished using “transfer learning”, where a model trained on one task is applied to another, or by using theory-driven models from relevant disciplines. Our own approach, described below, falls into the latter category, as we use social-psychological models to guide machine learning towards models that are both explanatory of data and predictive of relevant interactions.

The Dynamics Problem

Despite a widespread understanding that dynamic interactions between individuals are the core processes that need to be understood in order to make sense of small-group phenomena, much of classic group research in sociology and psychology has been surprisingly unaffected by the computational modeling techniques developed in other scientific disciplines that deal with dynamical systems. In contrast, artificial intelligence has a long history of attempting to model the behaviour of groups. For example, multi-agent systems (MAS) research aims to build teams of robots that can cooperate in working towards common goals, often by invoking strategic behaviours based on rational utility. This approach typically results in computationally complex strategic models that must account for many agents optimizing their utility functions simultaneously. On the other end of the spectrum, agents that replicate human behaviour are typically based on relatively simple models. Agent-based modeling (ABM) aims to replicate the emergent behaviour of groups using simple individual behaviours.

Multi-Agent Systems

A classic example of multi-agent systems research is the robot soccer "grand challenge" which has aimed (since 1998) to build a robot soccer team that can compete against the best human soccer teams (robocup.org). Other important application areas include teaching robots to conduct search and rescue and compete in multi-player computer games. As with much artificial intelligence research, most MAS work aims to build *rational* agents who are individual utility maximizers. That is, each agent has a set of preferences encoded in a utility function, which it uses to optimize its behaviour by computing expectations with respect to likely future scenarios. In order to enable group behaviours, such rational agents must model other agents' intelligent behaviour. If a cooperative solution is sought, agents may be endowed with a utility function that encodes group preferences. In a competitive situation, agents may need to plan for the worst-case strategic behaviour of the other agents. Further, a rational agent (call them "A") must concede that each other agent (call them "B") may also be optimizing in a similar way. Therefore, A must include a rational model for B, *as well as a nested model of B's model of A*. In fact, to be perfectly rational, agent A must continue to nest these models *ad infinitum*. Gmytrasiewicz and Doshi (2005) attempt to accomplish this, but their models are not scalable beyond a few agents in simple scenarios.

Rational choice in group behaviour has also been the subject of much investigation in economics and game theory. However, rationality leads to inconsistencies when considering simple games with social interdependence (e.g., social dilemmas). Humans in social dilemmas[1] are very good at finding what appear to be non-rational solutions that are more globally beneficial. Behavioural economists have tackled this problem by proposing a variety of mechanisms that explain the experimental evidence of prosocial (cooperative) behaviour in humans. Early work on motivational choice (Messick & McClintock, 1968)

[1] A social dilemma is a game with *uncompensated interdependencies* (externalities) (Kollock, 1998): each person's actions in the game affect other persons without their explicit consent (e.g., without compensating them).

proposed a probabilistic relationship between game outcomes (payoffs) and cooperative behaviour. This led to the proposition that humans make choices based on a modified utility function that includes some reward for fairness (Rabin, 1993) or penalty for inequity (Fehr & Schmidt, 1999). More recently, cooperative behaviour has been linked to altruism through factors like kinship, direct reciprocity, or indirect reciprocity via reputation (M. A. Nowak, 2006). However, it appears that fairness or inequity adjustments may not be comprehensive enough to account for human behaviour across all games, and a morality concept that is not based on outcomes provides a more parsimonious account (Capraro & Rand, 2017). The question of how this morality is defined is left as an open question, but it may be case that modeling human behaviour as motivated by emotion may exactly the type of social intuitionist (Haidt, 2001) model of moral and ethical reasoning that will explain some of these paradoxes.

Researchers have also found motivational and strategic solution concepts for cooperation based on group membership (Kollock, 1998). For example, Akerlof and Kranton have proposed an economic model in which an individual's utility function is dependent upon their identity (so called *identity economics*) (Akerlof & Kranton, 2000; Huettel & Kranton, 2012). Earlier work on *social identity theory* foreshadowed this economic model by noting that simply assigning group membership increases individual cooperation (Hogg, 2006; Tajfel & Turner, 1979). Other authors have confirmed that group membership influences individual choice (e.g., Charness, Rigotti, & Rustichini, 2007). This work has been contested with the counter-argument that group membership does not directly increase cooperation, but rather increases individuals' belief that others will cooperate (see Kollock, 1998). The difference is then between group membership as a motivational solution (being in a group actually changes one's payoff structure in some way), or as a strategic solution (being in a group changes ones beliefs about future events). In our recent work, we have shown that these two solution concepts may not be significantly different (Asghar & Hoey, 2015; Jung & Hoey, 2016, 2017). By considering

identity as a shared cultural and affective quantity, beliefs about group membership are directly connected to beliefs about strategic choices. That is, the very meaning of the group by definition is an affective one, and this affective sentiment is also explicitly connected to beliefs about behaviours (e.g., good people should do good things to good people and bad things to bad people). The forces of the resulting relational commitments of people to groups bear heavy weight upon the actions of group members (Lawler, Thye, & Yoon, 2009). We have shown that human behaviour in a social dilemma can be accounted for more closely using these basic principles (Jung & Hoey, 2016, 2017) and the mathematical structure of affect control theory (David R Heise, 2007, reviewed below).

In general, attempts to handle social interaction effects in artificial intelligence, game theory, and economics take the stance that the agent is still acting on rational and decision theoretic principles, but has a "modified" utility function (Bénabou & Tirole, 2006), with some "tuning" parameter that trades off social normative effects modeled as intrinsic rewards with the usual extrinsic rewards (e.g., exchange currencies). The tuning parameters are fit to data that is not accounted for by traditional economic models. Nevertheless, the fundamental problem persists in that an agent needs to optimize its behaviour by considering all possible strategic behaviours of other agents in order to compute a rational solution. These models lead to shallow (in time) and broad (in number of options considered) solutions due to limited processing power, and fail to provide convincing accounts of human social behaviour at a large scale.

Agent-Based Models and Social Simulation

Agent-based models (ABMs) are similar to multi-agent systems in that they consist of autonomous computational agents that interact with each other and thus generate an emergent, group-level outcome. However, in social science applications of ABMs, the goal is not to build a system that solves a problem (e.g., winning a soccer game) but to understand and explain the complex behaviours of a social system that are often not

trivially reducible to the properties of individual agents (for reviews, see Edmonds & Meyer, 2017; Smith & Conrey, 2007; Squazzoni, 2012). A primer is in order about the term "model", which is employed in a somewhat different sense than is typical in social psychology (as in "statistical model"). ABMs are mathematical implementations of theories in a piece of software about how agents process information internally and how they spread information to other agents. The software can then be used in a "social simulation", which is essentially a virtual experiment aimed at exploring the consequences of the chosen implementation of theory. Thus, ABMs are data-generating models. Of course, the data produced in social simulations can be analyzed with the same statistical methods as empirical data from observations with "real" agents/humans. In fact, social simulation modelers will often be interested in comparing the data patterns generated with a model to data patterns observed in the world. The logic behind this approach is: "if I can build a model that behaves like the real thing, I must have understood the real thing".

Many ABMs in social simulation have studied processes of attitude formation and diffusion in groups or societies, emphasizing the importance of social influence between agents (e.g., Deffuant, Neau, Amblard, & Weisbuch, 2000; Hegselmann & Krause, 2002; A. Nowak, Szamrej, & Latané, 1990). The agents in these models are usually very simple; e.g., an ABM might represent an agent's "opinion" as a number on a single dimension that changes according to a simple algebraic rule when subject to "influence" from another agent that is either a "neighbour" on a spatial grid or connected to the agent in a more or less realistic social network. For example, a homophilous agent may tend to change its behaviour to be more similar to the agents it interacts with most often (its "friends" and "co-workers"). The virtue of such models is that they show how even very simple mechanisms can produce, through the coupled interactions of many agents, complex group-level phenomena that are poorly understood – such as the contemporary ideological polarisation of Western democracies, for example (cf. Homer-Dixon et al., 2013). However, most ABMs lack even the most basic ingredients of intelligence (whether human or

artificial), namely the ability to reason, plan, and act in the world, let alone cooperate with other agents.

The psychological simplicity of agents in many social simulations has been the subject of much debate among modelers, who have always faced a tradeoff between principles of EROS (enhancing the realism of simulations) vs. KISS (keep it simple, stupid!) (cf. Jager, 2017). To put it polemically, it is not helpful to replace the dubious assumption of the full rationality of agents encountered in much AI research with the equally dubious assumption of full stupidity encountered in many ABMs. As the field matures and as computational resources become more ubiquitous, many social simulation researchers have moved to build more psychologically realistic ABMs. One example of such a model is GroupSimulator, developed by David R. Heise (2013). In this model, exchanges of behaviour among a group of computational agents are organized according to the structure of Bales' Interaction Process Analysis (Table 1). The choice of actions across IPA categories at each time step is computed according to the dynamic principles of affect control theory, which we review in more detail below as a possible starting point for a fruitful synthesis of AI and more traditional group-dynamics research.

Overcoming the Dynamics Problem

The work reviewed in this section lies at two modeling extremes. Multi-agent systems approaches attempt to build highly complex models of agent behaviour based on strategic analyses that are theoretically elegant and individually sensible, but fail to capture both the simplicity and emergent complexity of human group behaviour. Agent-based models, on the other hand, build simple descriptive models that are able to explain aggregate statistics of emergent human behaviour data, but often fail to account for individual interactions in specific settings. Our recent work bridges the gap between these two extremes, by proposing a dual-systems approach to artificial intelligence that combines rational reasoning with emotional motivations and attentional mechanisms. Our research

on affectively-motivated artificial intelligence is fundamentally different than models in MAS or ABM, in that it assumes the agent's reward is primarily extrinsic, but that attentional mechanisms based on affect control are used to focus on action choices that are aligned with the prevailing social order. The resulting solutions are therefore narrow (more focussed on socially aligned solutions) and deep, giving longer-term strategies of cooperation that are more predictive of human behaviour. While dual-systems approaches have also been investigated in beahvioural economics (Slovic, Finucane, Peters, & MacGregor, 2007), these typically relegate the affective system to a set of ad hoc heuristics that are descriptive of experimental human behaviours, but rarely grounded in social psychological theory. In the next section, we discuss our work in the development of psychologically grounded models, and show how they can be used to more parsimoniously account for human behaviour across a wider range of cooperative or competitive situations.

BayesACT: Integrating Social Psychological Theory and AI

Affect Control Theory

Affect control theory (ACT) is a mathematical theory that links social perception with identity, behaviour, and emotion in social interactions (for a comprehensive review, see Heise 2007). The theory draws on symbolic interactionism (Blumer, 1969; Neil J MacKinnon, 1994; Mead, 1934) as well as theories of psychological consistency (Heider, 1946; Simon & Holyoak, 2002) and cybernetic control (Powers, 1973; Robinson, 2007), proposing that people rely on linguistic representations with culturally-shared meanings to efficiently orient themselves within social interactions and anticipate the behavioural and emotional responses of others (David R Heise, 1979, 2007; Neil J MacKinnon, 1994; Smith-Lovin & Heise, 1988). Our motivation to maintain the cultural meanings associated with our own identities and the identities of others directly governs our interpersonal behaviours and emotions.

ACT uses the cultural meanings associated with labels for identities, behaviour, and

emotions to model how humans interpret and respond to social events. Based on classic work by Osgood and colleagues (e.g., Osgood, May, & Miron, 1975, 1957), the theory uses three universal semantic dimensions to measure cultural meanings for various concepts: 1) evaluation (good vs. bad), 2) potency (weak vs. strong), and 3) activity (calm vs. excited). Evaluation is associated with perceptions of warmth, likeability, and approachability. Potency is associated with perceptions of competence, dominance, and submission. Activity is associated with perceptions of social agency and action readiness (Rogers, Schröder, & Scholl, 2013; Scholl, 2013). Shared cultural knowledge, expressed on these dimensions (referred to collectively as EPA), describes and differentiates social concepts, with each concept possessing a specific pattern of affective meanings known as fundamental sentiments. Fundamental sentiments reflect how the members of a given culture view elements of the social world; they characterize how good, powerful, and active particular identities, behaviours, or emotions seem in general, outside of the context of social events. For example, we tend to see heroes as good, powerful, and active (2.6, 2.3, 2.1), mobsters as bad, powerful, and active (-1.2, 2.0, 1.2), senior citizens as good, powerless, and inactive (1.2, -0.0, -1.8), and dropouts as bad, powerless, and inactive (-1.7, -1.8, -1.5)[2].

Our fundamental sentiments for identities, behaviours, and emotions shift when they appear together in the context of social events. For example, a hero seems much more good, powerful, and active when he rescues a child (3.84, 1.96, 1.66) than when he compromises with a villain (.08, .92, -.13). These event-contextualized EPA meanings, known as transient impressions, capture the group's interpretation of actors, behaviours, and other elements of the situation and help to predict their behavioural and emotional responses to unfolding events. Affect control theory postulates that we can derive a group member's likely behavioural and emotional responses to a given situation from their transient impressions of that situation because human beings seek mental consistency between cultural expectations and social action (Heider, 1946; Schröder & Thagard, 2014;

[2]for historical reasons, EPA measurements are scaled to lie between -4.3 and +4.3

Simon & Holyoak, 2002). In other words, people act in ways that maintain the affective meanings associated with the group's interpretation of the situation, and expect others to do the same.

When our expectations about the identities and behaviours involved in an event are violated, we experience deflection, a sort of tension about the situation which signals that our experiences are out of alignment with cultural expectations. People seek to minimize deflection by acting in ways that maintain the group's interpretation of the situation; this is known as the affect control principle. Our social actions are planned and carried out to either maintain situational meanings or to bring them back into alignment with cultural expectations. Affect control theorists calculate deflection as the sum of the squared Euclidean distances between transient impressions of the identities and behaviours emerging from the situation and fundamental sentiments for these event elements. Thus, the lower the deflection, the greater the alignment between cultural expectations and situational circumstances. Deflection is much lower, for example, when a hero rescues a child than when they compromise with a villain.

Affect control theorists predict the emotions resulting from an event by solving for the EPA profile of the emotion that best explains the transient impressions experienced by the group given their fundamental sentiments. The theory thus asserts that our emotional response to a situation is determined both by our transient impressions of the event and the level of deflection the event produces. For example, nice events make us feel good. Events that are even better than we would expect based on the identities defining the situation make us feel even better. Researchers have found the ACT's predictions to accurately reflect the behaviours and emotions experienced in a variety of real-world social interactions (Freeland & Hoey, 2018; David R Heise & Lerner, 2006; Robinson & Smith-Lovin, 1992; Schröder & Thagard, 2013; Smith-Lovin & Douglas, 1992).

Recently, David R. Heise (2013) extended affect control theory to model small-group interactions by developing a simulation platform called GroupSimulator. Like the classic

ACT model of dyadic interactions on which it is based, GroupSimulator rests on the affect control principle, according to which agents strive to maintain the shared meanings of all identities involved in the interaction. The model capitalizes on the many strengths of ACT, such as its capacity to efficiently model the creative human interpretive process in a diversity of social situations, using a parsimonious dimensional structure to represent cultural meanings and small set of inputs to characterize events. In addition, compatibility between the theory's EPA measurement model and Bales (1999) SYMLOG measurement system (which is the basis of Interaction Process Analysis) allows for the classification of behaviours into IPA categories and facilitates the use of past groups data to validate simulation results.

GroupSimulator has been validated with mock jury data collected by Strodtbeck and Mann (1956). Participants in this study met in a judicial complex under the supervision of a court bailiff, listened to an audio recording of a trial, then deliberated to reach a judgement. Their deliberations were recorded and transcribed, and researchers manually classified the participants' interpersonal actions into IPA categories. David R. Heise (2013) was able to successfully reproduce the distribution of behaviours exhibited by the jurors in this study using GroupSimulator. The participants most frequently enacted task-related behaviours such as giving orientation and giving opinions, IPA categories 6 and 5 respectively. Second to these, participants most often engaged in socio-emotional behaviours such as agreeing and accommodating. The remaining five percent of the observed actions fall into one of the other nine IPA categories.

Yet, the model is not without its shortcomings. Although GroupSimulator is able to reproduce the behaviours of task groups, many of the parameters associated with social sense-making and turn-taking are external to the model rather than theoretically integrated components. Consequently, studies conducted with GroupSimulator are vulnerable to overfitting (creating an overly complex model to explain idiosyncrasies in the data). Two theoretical assumptions were also introduced in constructing this model to

address uncertainty in turn-taking: 1) the actor with the greatest deflection will act next when self-selection is possible; and 2) actors will choose to interact with the group member that will most effectively minimize their deflection. In addition, GroupSimulator features only one utility function, the minimization of deflection. In real-world small-group interactions, group members must balance identity maintenance with task-related priorities. The recent development of Bayesian affect control theory (*BayesAct*), provides a means to address many of these limitations.

Bayesian Generalization of ACT

The Bayesian generalization of ACT, called *BayesAct*, overcomes many of the limitations mentioned in the previous section (Hoey et al., 2016; Schröder et al., 2016). *BayesAct* adds three new elements to ACT, which can also be viewed as removing limiting assumptions of the theory.

1. *BayesAct* models all sentiments as probability distributions, thereby accounting for population-level differences in affective meanings for identities and behaviour that are likely replicated in personal uncertainties in social perception. Sentiment distributions can also be multi-modal, meaning that different viewpoints and multiple simultaneous identities and emotions of social agents are accounted for.

2. *BayesAct* includes a denotative state space that can represent other semantically meaningful elements of an interaction. Using this, utility can be defined beyond deflection to include other aspects of individual preference that are likely to affect agents' interpretations of and responses to events. *BayesAct* can therefore account for the tension involved in a social dilemma where individual and social gains are at odds with each other.

3. *BayesAct* allows for the simultaneous optimisation of all elements of an interaction, including identities, behaviours and turn-taking. The model seamlessly integrates

> notions of re-identification with behaviour alignment, and proposes a parsimonious account for how agents trade these off: as deflection gets too large to overcome with behaviour modifications, re-identification occurs and new modes of belief are introduced in the identity sentiments.

With these additions, *BayesAct* is constructed as a suitable basis model for task-oriented group interactions. *BayesAct* uses a probabilistic and decision theoretic model of stochastic control that arises in operations research called a partially observable Markov decision process (POMDP) (Åström, 1965). A POMDP is characterized by the maintenance of a *belief state* which is a distribution over the possible ways the world could be, and that represents everything an agent needs to know about its current state. A POMDP includes a utility function encoding agent preferences, and an optimization algorithm known as "dynamic programming" can be used to compute a mapping from belief states to actions of the agent. This mapping is called a *policy* and the policy that leads the agent to the highest utility (according to its preferences) is called the *optimal policy*.

An interesting property of a POMDP policy is that it may use "information gathering" actions. In the context of *BayesAct*, an agent can take actions that temporarily increase deflection in order to, for example, discover something about the interactant's identity, thereby helping the agent to decrease deflection in the long term, or to achieve some secondary reward. Information gathering is foundational to the reinforcement learning (RL) problem, in which an agent is confronted with a stream of experiences and rewards/punishments (positive/negative utilities according to its preferences), and must learn to optimize its behaviour, but must do so *while acting in the world* (Sutton & Barto, 2017). Such an agent is confronted with a catch-22: it must *explore* to find the best actions to take, but must also *exploit* its current knowledge of what works well.

In traditional RL, exploitation is seen as a cognitive skill requiring intense computation, since it involves predicting the future based on learned knowledge, and analyzing the costs and benefits of different strategies. Exploration, on the other hand, is

seen as something that could be guided by any number of (possibly affective) elements, such as being optimistic in the face of uncertainty. In *BayesAct*, the roles of exploration and exploitation are reversed. Exploitation consists of captializing on the learned socio-cultural knowledge of identity and behaviour dynamics to rapidly choose an affectively aligned action that promotes a social order. Exploration is now in the hands of the denotative reasoning engine that seeks actions nearby in the affective space (to the socially aligned action), but that may provide more individual reward. In *BayesAct*, the tradeoff between the two has a clear and simple meaning: it is a resource (time or energy) bound. If sufficient time or energy is available, then denotative "exploration" can occur, otherwise, connotative "exploitation" will rule the day.

BayesAct has been extended to take into account notions of the self (Hoey & Schröder, 2015), parallelling recent work on the affect control theory of self (Neil J. MacKinnon & Heise, 2010). Self-sentiments can be represented as distributions over the same affective space as identities and behaviour, and reflect persons' autobiographical memories about themselves as they really are. The affect control theory of self builds on the key insight of ACT, the affect control principle, showing that people are motivated to seek out situations that help them maintain their self-sentiments. The Bayesian affect control theory of self therefore includes a mechanism for selection of interactants into social situations (i.e., the alignment of self-sentiments with situational identity enactments), providing a theoretical justification for some of the seemingly ad hoc mechanisms used in GroupSimulator. A *BayesAct* version of GroupSimulator is under construction in order to allow us to carry out simulations. Some early examples can be found in Hoey and Schröder (2015).

Illustration: Group Dynamics in Virtual Software Development Teams

Understanding the social forces behind self-organized collaboration is increasingly important in today's society, where political problem-solving and the creation of economic

value occur less and less in formal, hierarchical organizations. Instead, we live in what scholars have described as an emerging distributed economy and digital democracy, where technological and social innovations are increasingly generated through informal processes of collaboration in and across startups, civic laboratories, fabrication labs and the like, often enabled through cheap and ubiquitous information and communication technology (e.g., see, Blowfield and Johnson (2013), Bogers and West (2012), Helbing and Pournaras (2015), Townsend (2013)). In the THEMIS.COG project [3], we study the open, collaborative development of software in online social coding communities like GitHub [4] as one key example of these economic changes. Exploring collaboration dynamics in communities like GitHub can further our understanding of the social and psychological mechanisms that drive the novel kind of human collaboration so central to the 21st century's economy and society.

Prior research suggests that people care at least as much about maintaining social relationships as they do about striving to maximize personal gains in their transactions with others. This makes intuitive sense, since maximizing one's gains depends on sustaining valuable relationships over time. Building on a long tradition of sociological theory and research, we hypothesize that identity dynamics explain how and why actors pursue each of these goals through interactions with others; our goal is to use Bayesian Affect Control Theory to predict and test collaborative dynamics. In this section, we first review the GitHub collaboration platform, then describe preliminary results from our work on solving the scaling problem and the dynamics problem. First, to tackle the scaling problem, we investigate automated methods for the analysis of IPA categories and sentiment from comment text on GitHub. Second, to tackle the dynamics problem, we use GroupSimulator to investigate some simple collaborative dynamics in simulation.

[3] themis-cog.ca
[4] github.com

Online Collaborative Networks

GitHub is a social coding platform where software developers from around the world come together to collaborate on software projects of common interest. The site enables software developers to work on the same software project (and even the same file in a project) simultaneously, and to merge their contributions without overwriting one another. The history of their contributions is saved, and one can always revert to an older version. While the vast majority of software projects on GitHub are open source, meaning that project-related code can be viewed, shared, and modified by other users and organizations on the site, this is not a requirement of the platform. There are indeed many closed source projects that use GitHub.

The following is a typical scenario in a GitHub project. A team of 10 developers is working on a software project. Alice and Bob are working on two different problems that both require the same file in the project to be edited. In a scenario without GitHub, one of them would have to wait till the other finishes and then make their changes. Using GitHub, both Alice and Bob can work simultaneously on local copies of the file. When they make their changes, they "commit" their contributions back to a central repository that maintains the project's history on the GitHub server. Let's say Alice commits her contribution first. Now when Bob tries tries to merge his contribution, GitHub will warn Bob that the file he is trying to make changes to has been changed since the last time he read it. Bob will have to carefully review the changes made so that he does not overwrite Alice's contribution.

Since such an infrastructure exists, a third person who is not part of the team can also make contributions to the project. Suppose John finds out there is a bug in the software, and he knows how to fix it. He can make a copy of the project repository in GitHub (called "forking" a project), then make a local copy of the repository on his local computer (called "cloning" the project). Once he makes changes to the project on his local computer, John can "push" his changes to the forked repository. Then he can create what

is called a "pull request" – a contribution that he is submitting to the project for review. A team member like Alice can look at John's contribution and either deem it acceptable or not. She can also initiate a discussion on this contribution and get the thoughts of others. They can ask John to make some changes and, when they deem it acceptable, they can "merge" the contribution to their project.

The above example is but one possible process by which software development may take place on GitHub. Each project team decides what process they are going to use for managing collaborative contributions. There are dozens of such process models. In order to collaborate on and contribute to a project, one has to follow the process outlined by the project team. Discussions and revisions by the group following the GitHub model is a crucial part of creating a relational meaning for the group of developers, which may later become a strong motivating force behind the group collaboration.

The Need to Go Further

Emotions and interaction processes play an important role in software collaborations. For example, emotions have been shown to affect task quality, productivity, creativity, group rapport and job satisfaction (De Choudhury & Counts, 2013). While positive emotions like happiness help people to be more creative, which is essential for successful software design (Fredickson, 2001), negative emotions such as fear can discourage developers from changing/refactoring their code (Ambler, 2002). While large-scale digital data traces for discussions of software projects are openly available through sites like GitHub, sentiment and emotional analysis can be challenging as affective content is embedded in technical discussions and punctuated with segments of code. Previous attempts include: Murgia, Tourani, Adams, and Ortu (2014), who perform a feasibility study of emotions mining using *Parrott's framework* on *Apache* issue reports; Guzman, Azócar, and Li (2014), who use lexical sentiment analysis to study emotions expressed in commit comments of open source projects; and Pletea, Vasilescu, and Serebrenik (2014b),

who use a Natural Language Text Processing tool (Natural Language Toolkit) to conduct a sentiment analysis of security-related discussions on GitHub. While studies like these yield interesting results, they are primarily descriptive in nature and therefore do not achieve the main purpose of performing sentiment analysis – to build models which are able to explain the behaviour of developers. There are several explanations for this limitation of the prior literature: either the techniques used are not robust enough, the dataset on which the analysis is performed is not large enough, or it is simply not possible to infer the behaviours of the developers from that dataset. Our work takes a hybrid approach, by leveraging social-psychological theory as a "top-down" model to guide the automated analysis. By grounding the analyses in affect control theory, we hope to show that a more generalizable model will be obtained.

Using Machine Learning to Code Group Interactions

In this section, we investigate methods for the automated analysis of both Interaction Process Analysis (IPA) categories (Bales, 1950) and emotion words. We investigate the 12 IPA categories shown in Table 1, as well as a set of ten emotions (positive: *Thanks, Calm, Cautious, Happy* and negative: *Sorry, Nervous, Careless, Aggressive, Defensive, Angry*). These emotion words were chosen to span ACT's three-dimensional emotional space (EPA) as identified by and related to IPA categories by David R. Heise (2013). In the following, we discuss our efforts towards classifying interactions on GitHub into these IPA and emotion categories. The ability to make such classifications will allow us to build group process simulations similarly to David R. Heise (2013), as reviewed in the following section.

We focus here on GitHub "pull" requests. We randomly selected 834 pull requests and a total of 3,000 pull request comments from GitHub in February 2017. Out of the 834 pull requests, 41 were open, 343 were closed without being merged, and 450 were merged. The comments were filtered to remove sections of code, then annotated by four people for the twelve IPA labels and ten emotions described above. One annotator was a co-author of

IPA group	**IPA Category**	**Example pull request comment**	**Emotions**
positive reactions	Shows solidarity	*im sure youll recover somehow*	Calm
	Shows tension release	*ooops sorry my mistake*	Sorry, Careless
	Agrees	*allright will do thanks for the feedback*	Thanks, Calm
attempted answers	Gives suggestion	*needs a metric tonne of docs*	Cautious
	Gives opinion	*love it*	Happy
	Gives orientation	*fucking hell im hungry now*	Aggressive, Angry
questions	Asks for orientation	*what if the file does not exist*	Nervous, Cautious
	Asks for opinion	*what about filtering by type and tag*	Cautious
	Asks for suggestion	*how could i show the name of the fighter that wins the turn*	Calm, Cautious
negative reactions	Disagrees	*for me just says linux which is not very useful at all*	Aggressive
	Shows tension	*um i dont know i dont remember changing that and probably did it by accident*	Nervous, Defensive
	Shows antagonism	*Kill this method with an axe and then burn its body*	Defensive, Aggressive

Table 2
IPA categories used in the study, along with example comments and emotion ratings.

the study, while three were hired on Amazon Mechanical Turk (MTurk). The three MTurk annotators had experience in programming and had heard of GitHub. Further, they were screened according to their ratings on an initial set of 50 pull request comments. Detailed instructions on how to annotate a particular pull request comment were provided, and each pull request comment could be annotated with a maximum of three IPA categories and a maximum of three emotions. The participants were also asked to filter out any unnecessary sections of code in the comment. More details can be found in (Rishi, 2017). Majority voting was used to threshold all the ratings; a comment was assigned an IPA/emotion label

if three out of four raters assigned it that label. Table 2 shows a few examples of the sentences corresponding to the various IPA and emotion categories.

Each word in each pull request was first mapped into a dimensional space defined by statistical patterns of words (so-called "word vectors") found in a large online corpus (the Google corpus). Word vectors were then weighted by their term-frequency, inverse document frequencies (TF-IDF) for each comment, which promotes words that are more important for the comment as a whole. Finally, a linear support vector machine (SVM) was trained on the resulting weighted vectors. An SVM optimizes a linear boundary between elements of two classes such that the two classes are maximally separated. Logistic regression, metric learning, and a variety of deep learning methods yielded similar results, see (Rishi, 2017) for details. We show the F1-scores (evenly weighted precision and recall) for a one-vs-all classification task of all IPA categories and all emotions in Table 3(a) and (b), resp. Parameters for the algorithms were set by searching exhaustively over a reasonably large, evenly spaced set of possibilities. Results are for a 5-fold cross-validation in which 4/5 of the data is used for training the SVM classifier, and 1/5 is used for testing, and this process is repeated for all five splits. From the results, it is clear that the task presents a significant challenge, which we believe can only be overcome by using more detailed emotional analysis of each comment [5]. We also examined aggregated IPA and emotion categories by grouping IPA categories into positive vs. negative reactions, and questions vs. attempted answers, and grouping emotions into positive and negative categories. The results in Table 3(c) show that this task is much simpler, and F1-scores over 0.85 can be achieved.

The results of our preliminary data analysis show that the task of sentiment and interaction analysis is a major challenge in cases with more objective conversations than what is usually attempted. And yet, it is known that subjective emotional and social interactions play a significant role in the online software development process. We have

[5] see (Alhothali & Hoey, 2015) for attempts in this direction

IPA Category	F1
Shows solidarity	56.8
Shows tension release	10.0
Agrees	**64.0**
Gives suggestion	33.4
Gives opinion	51.4
Gives orientation	58.6
Asks for orientation	36.2
Asks for opinion	22.9
Asks for suggestion	10.6
Disagrees	56.6
Shows tension	30.0
Shows antagonism	13.2

(a)

Emotion	F1
Thanks	54.7
Sorry	58.7
Calm	69.3
Nervous	23.6
Careless	15.7
Cautious	**69.8**
Aggressive	25.2
Defensive	16.7
Happy	2.5
Angry	0

(b)

Aggregated sets	F1-score
positive vs. negative reactions	86.6
questions vs. attempted answers	93.4
positive vs. negative emotions	98.5

(c)

Table 3

One vs. All classifications: (a) IPA categories; (b) Emotions; (c) aggregated classes

therefore exposed a significant gap for research in this area. Our current work is aimed at more fine-grained (sentence or word-level) sentiment analysis, and further group process analysis that may provide top-down information which can improve the overall effectiveness of the analysis. Longer term goals include the development of artificial agents to assist in software development by catalyzing more effective group processes online.

Theory Development and Generation of Hypotheses: Simulating Interactions on GitHub

We now turn to simulations of group interaction with GroupSimulator. We focus on two example simulations as an illustration of how GroupSimulator works, and the types of insights it can provide: 1) peer interactions occurring in a group of developers, and 2) interactions consisting of a leader and two newcomers. This allows us not only to compare a non-hierarchical group to a hierarchical one, but also to address a common and important type of interaction in online communities, the integration of newcomers (Marlow, Dabbish, & Herbsleb, 2013). We use GroupSimulator to examine the behaviours that are produced in each group, as well as who is enacting these behaviours and who is the target. We also

examine how experiences of deflection are distributed across members of the group.

In order to develop generative models of self-organized collaborations on GitHub, we first recruited a sample of 503 GitHub users and asked them to provide evaluation, potency, and activity ratings of 587 identities, behaviours, and other concepts (see Appendix A). We included concepts identified in the literature and by subject matter experts as salient features of GitHub interactions and tasks (Tsay, Dabbish, & Herbsleb, 2014), as well as those that are found in a large proportion of interactions on the site. Respondents were recruited through Qualtrics Panels, an organization that identifies and recruits potential study participants who meet specific eligibility criteria (e.g., demographic characteristics, expertise in a particular field). We oversampled with respect to both gender (50% female) and race (30% non-white). Participants ranged from eighteen to seventy-nine years of age, with most being in their thirties. While the majority of respondents had some college (17%) or a bachelor's (38%) or advanced degree (20%), others reported having some high school education (3%), a high school education (14%), or vocational training (5%).

Our simulation of a non-hierarchical group of developers consists of three good, powerful, and lively agents. The agents' identity sentiments are drawn randomly from a multivariate normal distribution, centered at 1.61, 1.91, and 1.76 in E, P, and A dimensions, respectively. In contrast, the leader and newcomer simulation consists of one very good, potent, and active identity (the group's leader), and two good but less potent and active identities (the two newcomers). Identity sentiments for the leader are drawn from a multivariate normal distribution centered at 2.67, 2.37, and 2.27, while the sentiments of the two newcomers are drawn from multivariate distributions centered at 1.78, .77, and .62. These values come from the survey described above.

Figure 2 displays the behaviour distributions predicted by GroupSimulator for a group of developers and a leader and two newcomers in the left and right panels, respectively. The x-axis indicates the IPA categories to which the behaviours were assigned. IPA categories refer to four clusters of behaviours: positive expressive behaviours

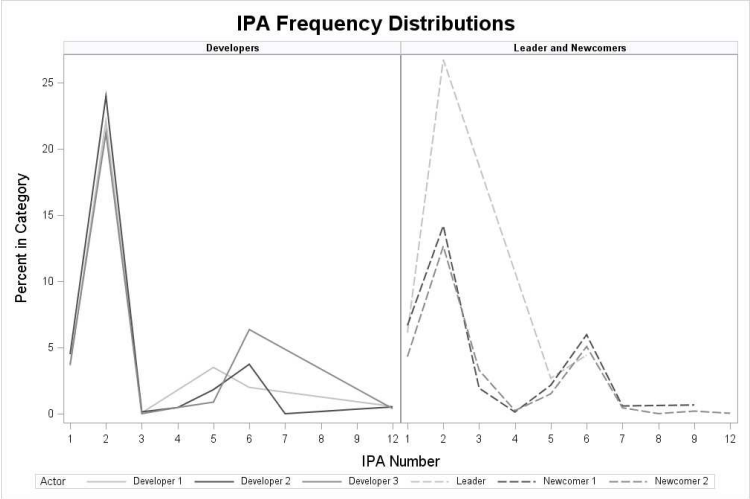

Figure 2. IPA Frequency Distributions for a Group of Three Developers and a Leader and Two Newcomers

(categories 1-3), behaviours associated with providing information or advice (categories 4-6), behaviours associated with soliciting information or advice (categories 7-9), and negative expressive behaviours (categories 10-12). The y-axis indicates the percentage of behaviours by IPA category each group member enacted, with the lines indicating the identities of each group member. The points indicate the frequency of behaviours in each category, with the absence of points indicating that the agent did not engage in a behaviour associated with that category. For example, developer 3 never agreed with other group members over the course of 1,500 turns across 220 simulations.

The difference between the frequency distribution of the group of developers and that of the leader and newcomers suggest that the simulation is able to capture the difference in the power dynamics implied by the identity labels. Although group members in both groups most frequently laughed or joked with others (category 2), the leader had many more opportunities than either developers or newcomers to engage in these behaviours. The leader also more frequently provided information than newcomers, and solicited for information or advice less often than newcomers. As expected, newcomers solicited for information and advice more often than either leaders or developers. Leaders also never engaged in negative expressive behaviours, most likely because the leader had numerous

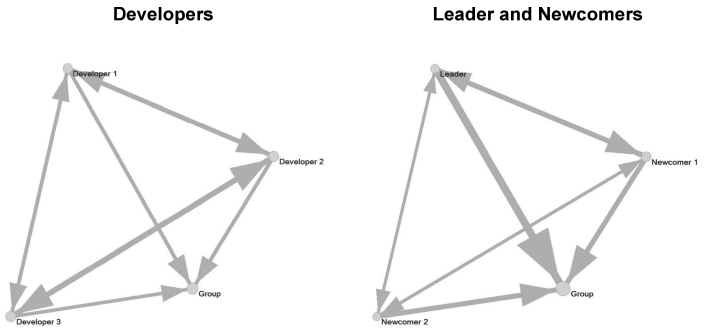

Figure 3. Interaction Networks of the Group of Three Developers and of the Leader and Two Newcomers.

opportunities to affirm its identity. The developers also fell into a superior/subordinate pattern, with developer 1 most often giving advice and suggestions while developers 2 and 3 most often solicited information and advice. The higher frequency of antagonism (category 12) among the developers, however, suggests that this was not always a happy arrangement.

Figure 3 clarifies these dynamics by showing the proportion of actions directed at each group member, and at the group as a whole. The left and right panels display the interaction networks of the developers, and of the leader and newcomers respectively. The nodes correspond to each group member and the group, and are sized by the number of behaviours directed at them. The arrows indicate behaviours directed by one group member at another member; the thickness of the arrows indicates the relative proportion of the total behaviours that occurred between each pair of actors. For example, the relatively equal weighting of the arrows directed by each developer towards the group indicates that each developer had essentially an equal number of opportunities to address the group, while the thicker arrows between developer 1 and developer 2 and between developer 2 and developer 3 indicate that interactions between these pairs of group members were more frequent than between the other group members. The roughly equal sizes of the nodes, however, indicate that each group member and the group were addressed by other group

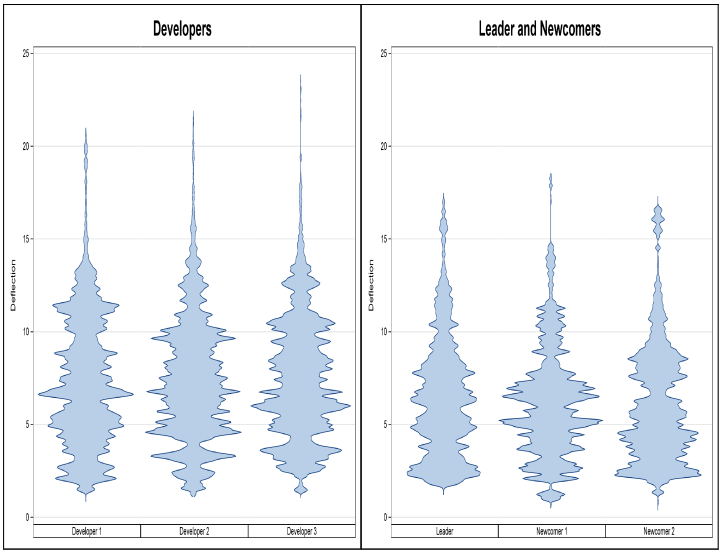

Figure 4. Deflection Experienced in a Group of Three Developers Compared to a Leader and Two Newcomers

members at nearly the same frequency.

The high proportion of behaviours enacted by the leader compared to newcomers and developers displayed in Figure 2 emerges from the interaction patterns featured in Figure 3. Leaders address the group, and newcomers tend solicit information and advice. The leader and newcomers tend to address the group more often than each other, with the exception of newcomer 1 and the leader. Nevertheless, the dominant interaction is the leader addressing the group, with far fewer pairwise interactions occurring than is true in the developers group.

Finally, Figure 4 displays the distribution of deflecting events experienced by each group member across the 220 runs, with the left and right panels referring to the developers and to the leader and newcomers respectively. The mean deflection and confidence interval of the developers is 7.41 (7.39-7.43). In contrast, the mean deflection and confidence interval of the leader and two newcomers is 6.39 (6.38-6.40) and 5.5 (5.51-5.62) respectively. The difference in the levels of deflection experienced by the developers compared to the leader and newcomers emerges from an interaction dynamic referred to by affect control theorists as the object diminishment effect. Being an object of

an interaction results in a loss of perceived potency, and thus is a source of deflection for potent identities such as developers (Smith-Lovin & Heise, 1988). Increased deflection results in group members directing compensatory behaviours meant to restore their perceived loss of potency to other group members which in turn leads to greater deflection, resulting in the pattern of peer-to-peer interactions shown in Figure 3 and the antagonism shown in Figure 2. In contrast, the lower relative potency of newcomers compared to the leader allows them to endure the leader's jokes and accept direction, and to direct most interactions towards the group rather than towards each other. By directing actions towards the group, the loss of perceived potency is distributed across the group reducing the tendency towards compensatory behaviours and thus reducing the overall level of deflection experienced by the group.

Discussion and Outlook: Next Steps in the Case Study

To summarize, affect control theory provides a well-grounded theoretical model that can make explicit predictions about interactions online in a collaborative group. These predictions are based on the notion of each group member holding an affective identity that is learnable, mathematically describable, and complementary to those of other group members (according to the principles of symbolic interactionist identity theories). These explicit predictions allow us to go much further and deeper than usual in data mining, as we can look for highly specific interactions of certain types in order to answer questions about the very nature of collaboration based on large-scale interaction traces in online collaboration networks. As an instantiation of ACT's theoretical model of group processes, augmented with the ability to model denotative state and multiple, conflicting identities, applications of *BayesAct* to interactions within online collaborative networks can provide direct empirical validation of group process theories, exposing novel areas of research and new social science questions.

General Discussion

The use of artificial intelligence (AI) in group process research has, to date, been somewhat limited. The limitations have primarily stemmed from difficulties researchers face in analyzing sufficient quantities of data on group interactions (the "scaling" problem), and from the inherent complexity of modeling human behaviours in groups (the "dynamics" problem). Artificial intelligence has largely been concerned with building systems that mine large data repositories in a somewhat "blind" fashion (i.e., failing to integrate prior theory, empirical evidence, or models of human interactions), and that build artificial agents based on a principle of rationality in the decision theoretic sense. These approaches fail to yield sufficiently rich or detailed models of human behaviour, especially within groups. In contrast, our work in building emotionally aligned AI is built upon a foundation of social-psychological theorizing about the role of emotion in group behaviour. Its fundamental tenet is that relational attachments between individuals and between individuals and groups define social orders that are strong, long lasting, and cooperative. These attachments form the basis for much human social interaction.

A key area of application for emotionally-aligned AI agents is online collaborative networks. More than ever, technological and social innovations are enabled by information and communication technologies and are generated through informal, distributed processes of collaboration, rather than in formal, hierarchical or market-based organizations. Although an individualization narrative pervades much theorizing about twenty-first century human interactions, an alternative socio-relational narrative has recently developed in which relational and affective person-to-group ties are understood as a keystone of networked coordination and effectiveness (Lawler et al., 2009). Relational ties grow from repeated interactions in groups with a shared responsibility in which positive emotions are created. Attribution by group members of their feelings to the group further strengthens the relational ties, creating a self-reinforcing mechanism for group coordination. Affect (emotion) is the essential element that fosters and promotes this strong group equilibrium.

Shared responsibility and positive affective interactions make the group salient and endow it with a moral and normative force upon the individual group members. Groups thus endowed are powerful agents for the mobilization of collaborative human efforts and collective action.

Our claim is that the foundations of human group behaviour in situations like these are likely to be based in a socio-affective mechanism that is socially transmitted and that encodes a social order (Hoey et al., 2016; Schröder et al., 2016). This is a radically new view for AI, as it starts from the premiss that humans are primarily *social* animals, rather than individualistic and rational ones. Strong and persistent ties in human networks are *relational* rather than *transactional* (Lawler et al., 2009). In this view, rationality exists at the level of groups of agents, not of individuals. Intelligence is defined by a social order that exists in a group and is internalised by each member through *affective dynamical structures* of *roles* or *identities*. Members of a group learn these structures as children, growing to assume a set of identities within the structures as adults. Members seek out other members of the group that play complementary roles, and enact a joint behaviour for their chosen relationship. Small-scale breakdowns are handled through a restorative set of multi-modal communicative cues that are displayed in the voice, face, gestures, and body, and are commonly referred to as *"emotion"* signals. Larger-scale breakdowns are handled by cognitive skill in creating new structures that are reified and internalized by group members (Berger & Luckmann, 1966). The dynamics of role relationships, coupled with human ability to cognitively explore in a time- and energy- bounded fashion, using reason and rationality, allow the entire group to build, maintain, enact, and transform a *social order* (Goffman, 1963) that is jointly optimal for survival. Our ability to build computational models of these processes based on the *BayesAct* framework allows us to bring the full weight of technological advances in AI to bear on the problem of how to model these processes in real networks involving thousands of humans and agents.

The presence of artificial agents in human social networks is growing. From chatbots

to robots, human experience in the developed world is moving towards a socio-technical system in which agents can be technological or biological artifacts, with increasingly blurred distinctions between. Our aim is to study computational models of affect and emotions from a social perspective in order to ensure that groups develop in this socio-technical world that are effective, efficient, moral and ethical, and that these groups positively reinforce basic human needs for strong, positive, and cohesive relationships. Our belief is that endowing artificial agents with socially aligned reasoning capabilities about affect is a foundational element in the construction of the socio-technical world that we are living in. Building on a long tradition of sociological theory and research, we propose that identity dynamics explain core motivations of actors in their interactions with others, and offer a mathematically precise model that can be used to predict and test collaborative dynamics. The general assumption is that humans are motivated in their social interactions by affective alignment: they strive for their social experiences to be coherent at a deep, emotional level with their sense of identity and general worldviews as constructed through culturally shared symbols. This affective alignment creates cohesive bonds between group members, and is instrumental for collaborations to become relational group commitments.

Acknowledgments This work was supported through the Trans-Atlantic Platform by NSERC and SSHRC (Canada), the DFG (Germany) and the NSF (United States). The GitHub analysis described was reviewed by and received full ethics clearance by the University of Waterloo Office of Research Ethics Review Board.

References

Akerlof, G. A. & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, *115*(3), 715–753. eprint: http://qje.oxfordjournals.org/content/115/3/715.full.pdf+html

Alhothali, A. & Hoey, J. (2015, May). Good news or bad news: using affect control theory to analyze readers' reaction towards news articles. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 1548–1558). Denver, Colorado: Association for Computational Linguistics.

Ambler, S. (2002). *Agile modeling: effective practices for extreme programming and the unified process*. John Wiley & Sons.

Asghar, N. & Hoey, J. (2015). Monte-Carlo planning for socially aligned agents using Bayesian affect control theory. In *Proc. uncertainty in artificial intelligence (UAI)* (pp. 72–81).

Åström, K. J. (1965). Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. App. 10*, 174–205.

Bales, R. F. (1950). *Interaction process analysis: a method for the study of small groups.* Addison-Wesley.

Bales, R. F. (1999). *Social interaction systems: theory and measurement*. New Brunswick, NJ: Transaction Publishers.

Bénabou, R. & Tirole, J. (2006). Incentives and prosocial behaviour. *American Economic Review*, *96*, 1652–1678.

Berger, P. L. & Luckmann, T. (1966). *The social construction of reality: a treatise in the sociology of knowledge*. Penguin.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blowfield, M. & Johnson, L. (2013). *Turnaround challenge: business and the city of the future*. OUP.

Blumer, H. (1969). *Symbolic interactionism: perspective and method*. Prentice-Hall.

Bogers, M. & West, J. (2012). Managing distributed innovation: strategic utilization of open and user innovation. *Creat. Innov. Manage. 21*, 61–75.

Capraro, V. & Rand, D. G. (2017, May 8). *Do the right thing: preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality*. SSRN.

Charness, G., Rigotti, L., & Rustichini, A. (2007). Individual behavior and group membership. *The American Economic Review*, *97*(4), 1340–1352.

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, *7*(3), 145–156.

De Choudhury, M. & Counts, S. (2013). Understanding affect in the workplace via social media. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 303–316). CSCW '13. San Antonio, Texas, USA: ACM.

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *3*(01n04), 87–98.

Edmonds, B. & Meyer, R. (2017). *Simulating social complexity: a handbook*. Springer.

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

Felbol, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017, October). *Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm*. arXiv:1708.00524v2.

Forsyth, D. R. (2018). *Group dynamics*. Cengage Learning.

Fredickson, B. (2001). The role of positive emotions in positive psychology. *American psychologist*, *56*(3), 218–226.

Freeland, R. & Hoey, J. (2018, April). The structure of deference: modeling occupational status using affect control theory. *American Sociological Review*, *83*(2).

Gmytrasiewicz, P. & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, *24*, 49–79.

Goffman, E. (1963). *Behavior in public places*. New York: The Free Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. http://www.deeplearningbook.org. MIT Press.

Guzman, E., Azócar, D., & Li, Y. (2014). Sentiment analysis of commit comments in github: an empirical study. In *Proceedings of the 11th working conference on mining software repositories* (pp. 352–355). MSR 2014. Hyderabad, India: ACM.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Hegselmann, R., Krause, U. et al. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, *5*(3).

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of psychology*, *21*(1), 107–112.

Heise, D. R. [David R]. (1979). *Understanding events: affect and the construction of social action*. CUP Archive.

Heise, D. R. [David R]. (2007). *Expressive order: confirming sentiments in social actions*. Springer.

Heise, D. R. [David R.]. (2013). Modeling interactions in small groups. *Social Psychology Quarterly*, *76*(1), 52–72.

Heise, D. R. [David R] & Lerner, S. J. (2006). Affect control in international interactions. *Social Forces*, *85*(2), 993–1010.

Helbing, D. & Pournaras, E. (2015). Build digital democracy. *Nature*, *527*, 33–34.

Hirschberg, J. & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266.

Hoey, J. & Schröder, T. (2015). Bayesian affect control theory of self. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 529–536).

Hoey, J., Schröder, T., & Alhothali, A. (2016). Affect control processes: intelligent affective interaction using a partially observable Markov decision process. *Artificial Intelligence*, *230*, 134–172.

Hogg, M. A. (2006). Social identity theory. In P. J. Burke (Ed.), *Contemporary social psychological theories* (Chap. 6, pp. 111–136). Stanford University Press.

Homer-Dixon, T., Maynard, J. L., Mildenberger, M., Milkoreit, M., Mock, S. J., Quilley, S., ... Thagard, P. (2013). A complex systems approach to the study of ideology: cognitive-affective structures and the dynamics of belief systems. *Journal of Social and Political Psychology*, *1*(1), 337–363.

Huettel, S. A. & Kranton, R. E. (2012). Identity economics and the brain: uncovering the mechanisms of social conflict. *Philosophical Transactions of the Royal Society B*, *367*, 680–691.

Jager, W. (2017). Enhancing the realism of simulation (EROS): on implementing and developing psychological theory in social simulation. *Journal of Artificial Societies and Social Simulation*, *20*(3).

Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Jung, J. D. & Hoey, J. (2016). Grounding social interaction with affective intelligence. In *Proceedings of the canadian conference on ai*. Victoria, BC.

Jung, J. D. & Hoey, J. (2017). *Socio-affective agents as models of human behaviour in the networked prisoner's dilemma*. http://arxiv.org/abs/1701.09112.

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, *83*(1), 126–137.

Kerr, N. L. & Tindale, S. (2014). Methods of small group research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 188–219). New York: Cambridge University Press.

Kollock, P. (1998). Social dilemmas: the anatomy of cooperation. *Annual Review of Sociology*, *24*, 183–214.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2018). The geometry of culture: analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Law, E., Gajos, K. Z., Wiggins, A., Gray, M. L., & Williams, A. (2017). Crowdsourcing as a tool for research: implications of uncertainty. In *Proc. of computer supported cooperative work (CSCW)*.

Lawler, E. J., Thye, S. R., & Yoon, J. (2009). *Social commitments in a depersonalized world*. Russell Sage Foundation.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... Gutmann, M., et al. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721–723.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

MacKinnon, N. J. [Neil J]. (1994). *Symbolic interactionism as affect control*. SUNY Press.

MacKinnon, N. J. [Neil J.] & Heise, D. R. [David R.]. (2010). *Self, identity and social institutions*. New York, NY: Palgrave and Macmillan.

Marlow, J., Dabbish, L., & Herbsleb, J. (2013). Impression formation in online peer production: activity traces and personal profiles in github. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 117–128). ACM.

Mead, G. H. (1934). *Mind, self and society*. Chicago University of Chicago Press.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Messick, D. M. & McClintock, C. G. (1968). Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology*, *4*, 1–25.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, *8*(5), e64417.

Murgia, A., Tourani, P., Adams, B., & Ortu, M. (2014). Do developers feel emotions? an exploratory analysis of emotions in software artifacts. In *Proceedings of the 11th working conference on mining software repositories* (pp. 262–271). MSR 2014. Hyderabad, India: ACM.

Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: a dynamic theory of social impact. *Psychological Review*, *97*(3), 362–376.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*, 1560–1563.

Osgood, C. E. (1962). Studies on the generality of affective meaning systems. *American Psychologist*, *17*(1), 10.

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. University of Illinois Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning. 1957*. Urbana: University of Illinois Press.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135.

Pletea, D., Vasilescu, B., & Serebrenik, A. (2014a). Security and emotion: sentiment analysis of security discussions on github. In *Proc. of mining software repositories (MSR)*.

Pletea, D., Vasilescu, B., & Serebrenik, A. (2014b). Security and emotion: sentiment analysis of security discussions on github. In *Proceedings of the 11th working conference on mining software repositories* (pp. 348–351). ACM.

Powers, W. T. (1973). *Behavior: the control of perception*. Aldine.

Rabin, M. (1993). A theory of fairness, competition and cooperation. *The American Economic Review*, *83*(5), 1281–1302.

Rishi, D. (2017). *Affective sentiment and emotional analysis of pull request comments on github* (Master's thesis, University of Waterloo).

Robinson, D. T. (2007). Control theories in sociology. *Annual Review of Sociology*, *33*.

Robinson, D. T. & Smith-Lovin, L. (1992). Selective interaction as a strategy for identity maintenance: an affect control model. *Social Psychology Quarterly*, 12–28.

Rogers, K. B., Schröder, T., & Scholl, W. (2013). The affective structure of stereotype content: behavior and emotion in intergroup context. *Social Psychology Quarterly*, *76*(2), 125–150.

Savage, M. & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885–899.

Schermuly, C. C. & Schölmerich, F. (2017). Analyse von gruppen in organisationen. In S. Liebig, W. Matiaske, & S. Rosenbohm (Eds.), *Handbuch Empirische Organisationsforschung* (pp. 491–512). Springer.

Schermuly, C. C., Schröder, T., Nachtwei, J., & Scholl, W. (2010). The discussion coding system (DCS): a valid and economical instrument to code interactions in organizations. *Zeitschrift Für Arbeits- und Organisationspsychologie*, *54*(4), 149–170.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public Opinion Quarterly*, *80*(1), 180–211.

Scholl, W. (2013). The socio-emotional basis of human interaction and communication: how we construct our social world. *Social Science Information*, *52*(1), 3–33.

Schröder, T., Hoey, J., & Rogers, K. B. (2016). Modeling dynamic identities and uncertainty in social interaction: Bayesian affect control theory. *American Sociological Review*, *81*, 828–855.

Schröder, T. & Thagard, P. (2013). The affective meanings of automatic social behaviors: three mechanisms that explain priming. *Psychological Review*, *120*(1), 255–280.

Schröder, T. & Thagard, P. (2014). Priming: constraint satisfaction and interactive competition. *Social Cognition*, *32*, 152–167.

Simon, D. & Holyoak, K. J. (2002). Structural dynamics of cognition: from consistency theories to constraint satisfaction. *Personality and social psychology review*, *6*(4), 283–294.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operations Research*, *177*(3), 1333–1352.

Smith, E. R. & Conrey, F. R. (2007). Agent-based modeling: a new approach for theory building in social psychology. *Personality and social psychology review*, *11*(1), 87–104.

Smith-Lovin, L. & Douglas, W. (1992). An affect control analysis of two religious subcultures. In V. Grecas (Ed.), *Social perspectives on emotion* (pp. 217–247). JAI.

Smith-Lovin, L. & Heise, D. R. [David R]. (1988). Affect control theory: research advances. *New York: Gordon and Breach*.

Squazzoni, F. (2012). *Agent-based computational sociology*. John Wiley & Sons.

Strodtbeck, F. L. & Mann, R. D. (1956). Sex role differentiation in jury deliberations. *Sociometry*, *19*(1), 3–11.

Sutton, R. S. & Barto, A. G. (2017). *Reinforcement learning: an introduction* (2nd ed.). MIT Press.

Tajfel, H. & Turner, J. C. (1979). An integrative theory of intergroup conflict. In S. Worchel & W. Austin (Eds.), *The social psychology of intergroup relations*. Monterey, CA: Brooks/Cole.

Townsend, A. (2013). *Smart cities: big data, civic hackers, and the quest for a new utopia*. Norton.

Tsay, J., Dabbish, L., & Herbsleb, J. (2014). Let's talk about it: evaluating contributions through discussion in github. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering* (pp. 144–154). ACM.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: what 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, *10*(1), 178–185.

Vallacher, R. R., Read, S. J., & Nowak, A. (2017). *Computational social psychology*. Routledge.

Wang, Y. & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257.