Toward a computational theory of social groups: A finite set of cognitive primitives for representing any and all social groups in the context of conflict

David Pietraszewski

Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
davidpietraszewski@gmail.com | https://www.mpib-berlin.mpg.de/en/staff/david-pietraszewski

Target Article

Keywords: Commonsense reasoning; computational theory; information-processing; intergroup relations; social cognition; social groups; triadic interaction; violence/aggression

The response made by David Pietraszewski to the commentary following this article can be found here.

Abstract

We don't yet have adequate theories of what the human mind is representing when it represents a social group. Worse still, many people think we do. This mistaken belief is a consequence of the state of play: Until now, researchers have relied on their own intuitions to link up the concept *social group* on the one hand and the results of particular studies or models on the other. While necessary, this reliance on intuition has been purchased at a considerable cost. When looked at soberly, existing theories of social groups are either (i) literal, but not remotely adequate (such as models built atop economic games), or (ii) simply metaphorical (typically a subsumption or containment metaphor). Intuition is filling in the gaps of an explicit theory. This paper presents a computational theory of what, literally, a group representation is in the context of conflict: It is the assignment of agents to specific roles within a small number of triadic interaction types. This "mental definition" of a group paves the way for a *computational theory of social groups* – in that it provides a theory of what exactly the information-processing problem of representing and reasoning about a group is. For psychologists, this paper offers a different way to conceptualize and study groups, and suggests that a non-tautological definition of a social group is possible. For cognitive scientists, this paper provides a computational benchmark against which natural and artificial intelligences can be held.

1. A hard problem: What is a group?

What is a group? More than a century into the study of groups and group identities, this question continues to vex. Human intuition suggests that groups exist (i.e., as far as we know, all typically developing humans on earth hold the mental concept of a social "group" as part of their perception of the world – it is a universal *emic* concept; Brown, 1991). Research furthermore shows that people hold reliable and consistent intuitions about what is and is not a group (e.g., Hamilton, Sherman, & Castelli, 2002). Thousands of scientific papers, hundreds of books (e.g., Brown, 2000; Forsyth, 2014/2019; Sedikides, Schopler, & Insko, 1998), and dozens of scientific journals (e.g., *Advances in Group Processes, Group Dynamics, Group Processes & Intergroup Relations,* etc.) are predicated on the existence of this construct. Yet what exactly are we talking about?

Almost all definitions agree that "group" means something more than just an aggregate and more than just a dimension of similarity out in the world (Campbell, 1958; Forsyth, 2014/ 2019; Ip, Chiu, & Wan, 2006; Wilder & Simon, 1998). A pile of leaves, for example, is not a group. A group is something more psychological, more agentic. Agency on its own, however, does not seem to be sufficient; people waiting for a bus are classically not a group, for example (Lickel et al., 2000). It seems then, that it is not enough that the entities are agents. There must also be coordination and some psychological representation among the individuals that they are so coordinated (Deutsch, 1962; Sherif, Harvey, White, Hood, & Sherif, 1961; Tajfel, 1982).

Thus, in more sophisticated definitions, we find a "relationship" (Lewin, 1948), "mutual awareness and interaction" (McGrath, 1984), "obligations," "influence" (Shaw, 1981), "interdependence" (Lewin, 1948), and "social cohesion" existing between individuals – based on "a common identity" that people "define themselves as members of," that "is recognized by at least one other" (Brown, 2000) and that people "care about" based on shared "values," "experiences," "interests," "kinship," and so on (see Forsyth, 2014/2019, for a review).

Undoubtedly, these definitions bring us closer to an adequate definition. The concept "group" of course refers to a very broad set of things. So a satisfactory definition – if it is to be inclusive – must be correspondingly broad and short on details. As Allport famously put it, “It is difficult to define an in-group precisely. Perhaps the best that can be done is to say that members of an in-group all use the term *we* with the same essential significance” (Allport, 1954/1958, p. 30).

Nevertheless, we should not yet be fully satisfied with these current definitions. The problem – correct as they are – is that by relying on one or more additional psychological constructs, these definitions are still not yet grounded in concrete, operationalizable terms. For instance, imagine that an artificial intelligence (AI) engineer approaches us and wants to build a machine intelligence capable of “groups.” What exactly would we tell them? Groups involve the formation of “expectations,” “obligations,” and a meaningful “we.” But expectations and obligations about what, exactly? And what exactly is the essential significance of *we*? We might answer with obligations to “support one another,” to “come to one another’s aid,” and so on. But this kind of explication still appeals to psychological concepts – all of which simply shift the goal posts. Already having an intuition or folk-psychological conception of groups is necessary to understand (and implement) what is meant by “support,” “aid,” “interdependence,” “we,” and so on. We are left then with an infinite regress of appeals to intuition. We know a group when we see it, in other words, but we couldn’t say exactly what it is. Or, in the more precise language of the type/token distinction (Peirce, 1931–1935): We can recognize *tokens* (particular instances) of groups, but we don’t have an explicit notion of what constitutes the *type* (the class itself).

In fact, there have been roughly two different ways of conceptualizing and studying groups in the behavioral sciences. The first seeks to capture how people intuitively represent and reason about groups, largely borne out of the ground-breaking theorizing of mid-twentieth century social psychology. The problem with this approach – as we have just seen – is that a necessary and sufficient definition of a group has not been forthcoming, and the definitions that do exist tend to remain rather psychological, if not verging on tautological.

A second approach has been to not rely on vague, psychological terms at all, but rather to operationalize the concept of a group as something concrete enough that it can be modeled or studied – which typically involves describing objective behaviors among multiple agents. This formalized study of n-person dynamics has been underway for several decades now – originating from a diverse set of fields including AI (Sandholm, Larson, Andersson, Sherry, & Tohmé, 1999), evolutionary biology (Koykka & Wild, 2017; Rusch & Gavrilets, 2020), the evolutionary social sciences (Böhm, Rusch, & Baron, 2018; Glowacki, Wilson, & Wrangham, 2017), primatology (Harcourt & de Waal, 1992), international relations (Schelling, 1966), and economics and decision-making (Schelling, 1956). These n-person dynamics have been studied in the lab (e.g., Bornstein, 2003; Gonzalez, Ben-Asher, Martin, & Dutt, 2015; Yamagishi, Jin, & Kiyonari, 1999) and in the field (Bissonnette et al., 2015; De Dreu & Van Vianen, 2001), and have been informed by a decades-long enterprise of game theoretic and optimality analyses (e.g., Burani & Zwicker, 2003; Gamson, 1961; Mesterton-Gibbons et al., 2011; Shehory & Kraus, 1998).

The problem with this second approach, however, is that these concrete implementations have not informed the issue raised by the first approach: namely, what – literally – is being represented in the mind when people are perceiving and reasoning about groups? For the purposes of defining a group, researchers in these areas are typically less concerned with directly assessing people’s intuitions of what constitutes a group (as in the first approach), and so simply side-step this issue and declare by fiat that a particular set of behaviors, decisions, or processes constitutes a “group” for their particular purposes (i.e., they posit – either implicitly or explicitly – an operational definition). This declare-by-fiat strategy has afforded great progress, as it allows researchers to use mechanistic techniques (e.g., using agent-based or analytic models; e.g., Goldstone, Roberts, & Gureckis, 2008; Gross & De Dreu, 2019).[1]

Nevertheless, the problem of defining a group sits awkwardly with respect to this second approach, as well. Namely, if the problem of defining a group were simply a matter of rendering the notion of a “group” into something concrete, then this second approach leaves us with an embarrassment of riches: There are dozens, if not hundreds, of concrete operationalizations of the concept “group” within this literature – including behaviors in the lab, the field, and the computerized world of agent-based models. Yet none of these seem to feed back onto the intuitive concept that they instantiate, such that we could point to any one concrete implementation or model and say “this is a group.” Or, more accurately, we would have to say that every single one is an instance of a “group” (at least to the intuition of the researcher and the reviewers who allowed the paper to be published as a study on “groups”) – all of which suggests that this second approach relies just as much on the human intuition of what constitutes a group as the first (albeit in a different way). In particular, intuition is being used to denote particular behaviors or dynamics as being about groups.

Consequently, as a scientific community, we have created a vast universe of group *tokens* – either described psychologically or operationalized concretely – without having a clear idea of what constitutes the *type*. Moreover, we now have the additional problem of not having a clear diagnosis of the original problem: that the problem of defining a group cannot simply be a matter of rendering the concept into something concrete. If it were, the second, concrete-operationalization approach would have already solved the problem of “What is a group?” long ago.

DAVID PIETRASZEWSKI is a Research Scientist at the Max Planck Institute for Human Development in Berlin, Germany. His work focuses on the psychological mechanisms for dealing with multi-agent coordination, cooperation, and conflict across the lifespan, and has been published in outlets such as *Proceedings of the Royal Society, Perspectives on Psychological Science, Cognition,* and *Leadership Quarterly*. He also addresses field-wide issues related to cognition and philosophy of the mind.

2. A solution: A computational theory

What is missing from current approaches to the study of groups – and what will finally resolve the issue of what is a group – is to have what the vision scientist David Marr (1982) referred to as *computational theories*: explicit information-processing theories of what the mechanisms that make up the mind are representing and operating upon when, pre-theoretically, one is reasoning about some phenomenon (which in this case would be “groups”).[2] Computational theories resolve the long-standing tension between the two approaches described above by (1) explicitly acknowledging the importance of psychological representation –

as in the first approach (i.e., they do not naively place the notion of a "group" out in the world, but rather, as a representation within the mind), while at the same time (2) grounding that psychological representation in something concrete, objective, and non-tautological – as in the second approach.

Computational theories highlight the issue of *computational adequacy*: that if what one has stipulated as going on within some information-processing entity is adequate for creating the phenomenon of interest, without intervention from a god-like agent or intelligence (Chomsky, 1980; Marr, 1982; Minsky, 1961, 1974; Pietraszewski & Wertz, 2021; Tooby & Cosmides, 1992). This issue of computational adequacy differentiates the present approach from what we have described as the second, concrete-operationalization approach to the study of groups. Nearly all of that work takes as its initial starting point the imposition of some sort of predefined structure onto the world – typically a game structure, such as the prisoner's dilemma game (or kindred snow-drift game; see Axelrod, 1984; Oliver, 1993; Perc & Szolnoki, 2010; Vainstein, Silva, & Arenzon, 2007; Worden & Levin, 2007, for reviews). These and other predefined structures render the enterprise tractable, allowing researchers to not have to worry about certain pesky details.

In contrast, on our computational-theoretic approach, we are not allowed to think that the world can come preformed into games or structure. Instead, the representational system itself must impose some sort of framing onto the blooming, buzzing confusion of the real world. That is, the representational system has the job of imposing its own "game" structure onto the real world, and then "playing" an internally represented game, by using cues in the real world to run its simulations. As scientists, we then face the same problem as that faced by the mind: determining what – out of all the things that humans are doing out in the real world – constitute "group" things. Or, in the parlance of AI and philosophy, we are forced to confront *the relevance problem* (e.g., Dennett, 1988).

Colloquially, then, our computational-theoretic approach can be thought of in the following way: We are engineers, and our job is to build a robot that can interface with the real world. The question, "What is a group?", then becomes a question about what could in principle be a group representation be in such a robot. That is, how does our robot "see" groups? What is the definition of a group within its software?[3]

The proposed computational theory is as follows: The folk-psychological construct "group" corresponds to changes within a framework of event types (what computer scientists and AI researchers would refer to as an *event grammar* or *event calculus*; e.g., Mueller, 2015). Within this framework, a group representation corresponds to the probabilistic assignment of agents to particular roles within those event types.

Here, we will consider event types featuring cost impositions. We are thereby addressing what a "group" representation is in the context of conflict. Doing so is not meant to suggest that there are *only* conflict-based group representations within the mind. Rather, we simply have a plausible account for conflict – which is no trivial matter, as our account must describe what is universally true of all instances of group membership in all instances of real or imagined conflict, all the while describing that representation in its most compressed, reduced form. At the end of the paper, we will speculate as to how to think about other, non-conflict-based event types, and thereby make progress toward a complete computational theory of social groups.

The remainder of the paper will explain the proposed conflict-based event-type computational theory of a group representation – all in a conceptual, non-technical manner – and will try to convey a sense of what kinds of research questions it opens up, all while attempting to head-off a number of likely misunderstandings.

3. Starting with the reduction problem

Any computational theory must first deal with what is called the *reduction problem* in AI: how to take an apparently unbounded, infinite phenomenon and reduce it to a finite, solvable set (Russell & Norvig, 1995). For example, vision researchers must postulate a finite set of cognitive primitives that allow a viewer to identify and recognize any and all objects (e.g., Marr & Nishihara, 1978). Analogously for groups, we must postulate a finite number of cognitive primitives that allow all tokens of group conflict – from playground fights to international conflict, and everything else in between – to be represented and thereby operated upon within the mind.

A good conceptual starting point for solving the reduction problem for group representation is to start with a broader reduction problem: that of multi-agent (or n-person) conflict, which refers to when more than two agents come into conflict with one another. For decades, the complexity and near-infinite variety of n-person conflict has posed a problem: How can a finite set of mental representations and decision-rules handle the complex dynamics that arise when more than two agents come into conflict with one another (e.g., Byrne & Whiten, 1988; Harcourt & deWall, 1992; Sherratt & Mesterton-Gibbons, 2013)? How, in other words, can the mind "see" all instances of multiperson conflict – of which group-related phenomenon are but a subset?

A solution to this problem was offered recently (Pietraszewski, 2016), and while a solution to this broader reduction problem of n-person conflict does not solve the specific problem of what a group representation is, it does narrow down the scope of possible solutions – in the same way that knowing the answer to the question "Is it bigger than a bread-box?" narrows down the scope of possible solutions in the game of 20 questions. This n-person solution was discovered by asking the question, "What would always be true of n-person conflict over evolutionary time?"[4] From the multigenerational perspective of evolutionary time, the particular details of conflict-related events will always be different. Who is involved, why, and the nature of the conflict itself will not be stable across multiple generations. Therefore, psychological processes for handling conflict cannot be too-tightly tied to any of these particular details, but rather must operate over a class of abstracted invariances (Jackendoff, 1992). Once we strip these ever-changing details away, what remains are the following principles:

- *Conflict can be understood as the contingent delivery of costs.* An inclusive definition conceptualizes conflict as the contingent delivery of costs between agents; an expectation or realization of the contingency *if you do X, I will do Y* (Archer, 1988; Campbell, 2015; Daly, 2015; Hardy & Briffa, 2013). This conceptualization accommodates exchanging costs in kind – exchanging an eye for an eye and a tooth for a tooth – and accommodates an exchange of different costs. For example, when a child pushes another for stealing a toy, they are exchanging one cost for another, repaying lost enjoyment with physical harm.
- *N-person conflict can be decomposed into triadic – but not dyadic – interactions*. What sets multi-agent or group conflict

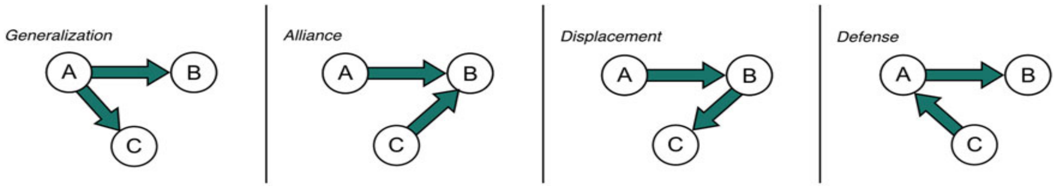


Figure 1. The four ways in which a third party (C) can become involved in an otherwise dyadic conflict between A and B. From left to right, A's attack on B is followed by: A also attacking C; C attacking B; B attacking C; C attacking A. These are *Generalization*, *Alliance*, *Displacement*, and *Defense*, respectively.

apart from dyadic conflict is the involvement of a third party. For example, if A attacks B, and then in response B attacks A, and so on and so forth, this may become a protracted conflict, but it will only ever become a multi-agent conflict if at least one third party, C, becomes involved (Caplow, 1959; Grammer, 1992; Harcourt, 1988; Harcourt & deWall, 1992; Heider, 1958; Liska, 1962; Patton, 1996, 2000; Pietraszewski, 2012, 2016; Strayer & Noel, 1986; Von Neumann & Morgenstern, 1944).

- *There are only four ways that a third party can become drawn into a conflict within these triads*. If A attacks B, and a third party, C, becomes involved, there are only four possible ways in which this might happen (Chase, 1985; Strayer & Noel, 1986), depicted in Figure 1:

1. *Generalization*: A attacks B, then A attacks C
2. *Alliance*: A attacks B, then C attacks B
3. *Displacement*: A attacks B, then B attacks C
4. *Defense*: A attacks B, then C attacks A

These triadic interactions characterize what is always true of n-person conflict. They are the smallest set of "building block" units that can be strung together over time to describe an infinite number of conflict dynamics – similar to the way a finite number of letters can generate an infinite number of linguistic meanings. Past work had used these four triadic interactions as a coding scheme for documenting real-world multi-agent conflict (Chase, 1985; Strayer & Noel, 1986). However, these triadic interaction types – because they describe all of the ways in which a third party may become involved in a conflict – may also be the computational building blocks out of which all n-person conflicts are cognitively represented (Byrne & Whiten, 1988; Harcourt & deWall, 1992; Sherratt & Mesterton-Gibbons, 2013). We will refer to these as *triadic primitives*.

Crucial to this proposal is that (1) any one agent can occupy each of the 12 different roles over time (i.e., A, B, and C are representational slots into which particular agents are inserted; they are not agents themselves) and (2) that these triadic interactions will string together or concatenate over time as interactions unfold. Consequently, the human mind need not represent an infinite number of conflict dynamics: it need only reason about each of these four interaction types, and then string them together over time as real or hypothetical events unfold. Such an architecture – described at this high level of abstraction of strings of concatenated triadic primitives – solves the reduction problem with respect to n-person conflict, allowing a finite cognitive architecture to represent an infinite variety of conflict dynamics (see Pietraszewski, 2016, for details).

4. The proposed solution: The mind's definition of a "group" in the context of conflict

With the broader n-person reduction problem tentatively addressed, we can turn back to the original problem of characterizing the mind's definition of group in the context of conflict: If the above four triadic primitives describe all multi-agent conflict, then it would seem likely that the concept of "group" is somehow present in these triadic primitives. But how? Strings of triadic primitives do not yet map onto or produce the intuitive concept "group." How do we go from these triadic primitives to the concept of "group"?

The proposed solution is as follows: *a group (in the context of conflict) is a set of specific roles within these four triadic primitives*, where *role* simply refers to one of the three interactant-slots within each triadic primitive (e.g., being A in *Generalization*). These roles are depicted in Figure 2:

What it means that individuals are members of a group, then, is that:

(i) in *generalization*, they are in roles B and C
(ii) in *alliance*, they are in roles A and C
(iii) in *defense*, they are in roles B and C
(iv) in *displacement*, they are in roles A and C

These *group-constitutive* roles (again, with roles referring to the agent and/or patient roles taken within the triadic primitives)

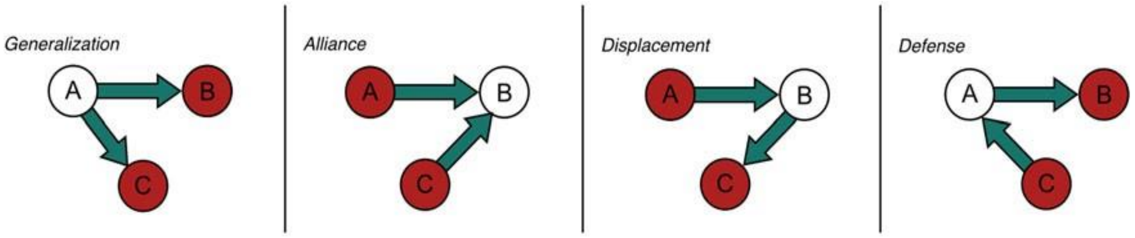


Figure 2. The roles that define the mental concept "group" in the context of conflict (depicted in red). These we swill call *group-constitutive roles*.

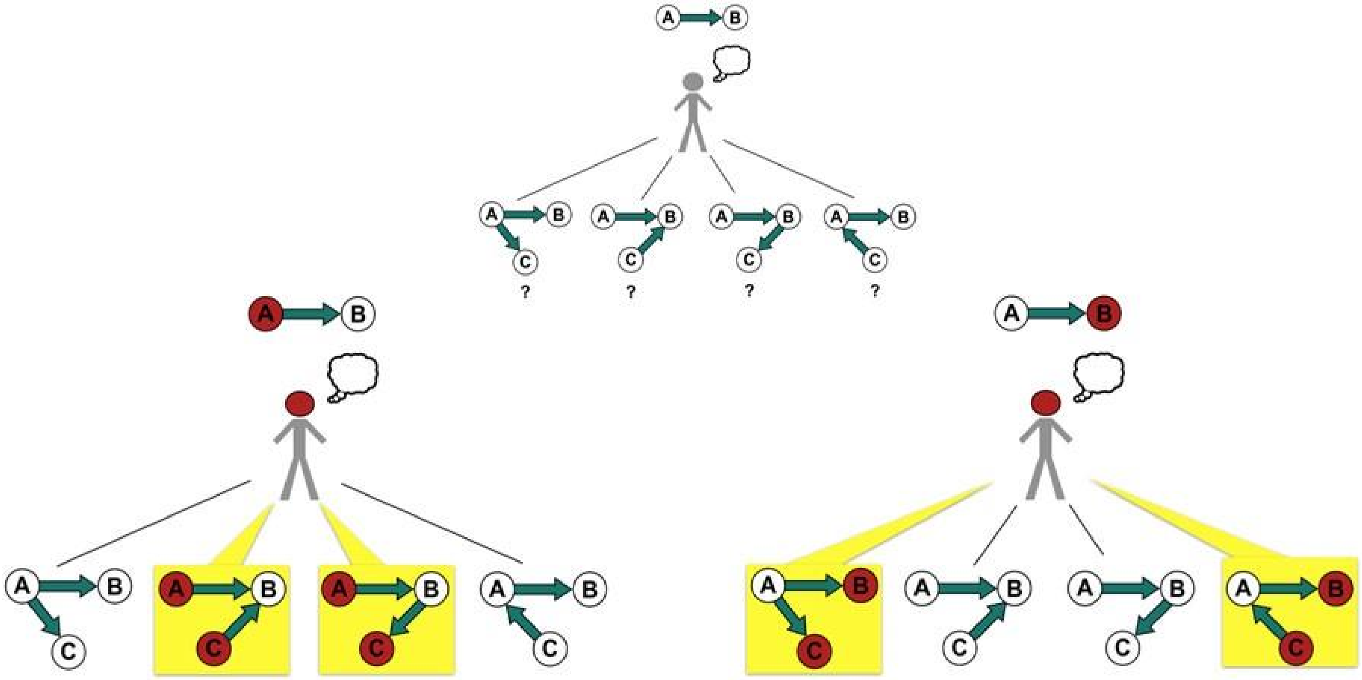


Figure 3. A single agent (depicted as the stylized figure) observes a cost imposition occurring between A and B (top panel). Without knowing the group memberships of A, B, and the agent, we have no basis for anticipating which, if any, of the four possible triadic interactions are likely to happen as a consequence of A attacking B. However, once the group membership of the three agents is marked (group membership is denoted in red in the bottom two panels), certain triadic interactions become more likely to occur (those that are highlighted), whereas others become less likely to occur (those that are not highlighted).

ground "choosing one agent over another" and "taking sides" in a conflict into a finite set of mental representations. The precise, casual claims are as follows:

- Cost/benefit calculations within each individual agent produce the above triadic interactions.
- Certain classes of cost/benefit calculations will cause certain agents to occupy the above roles with respect to one another.
- This class of cost/benefit calculations "hang together" over evolutionary time. That is, the relationships between agents that cause them to occupy the above roles with respect to one another constitute an enduring entity over evolutionary time.
- This enduring entity allows for selection to create a summary representation in the mind. This is a mental entry for *agents who are in the kind of relationship which causes them to occupy these particular roles with respect to one another.*
- That summary representation – precisely, the group-constitutive roles within the four triadic primitives – is our mental concept "group" in the context of conflict.

The claim, then, is that on-the-ground relationships will cause agents to occupy these roles with respect to each other. That there are such relationships creates a conceptual entry or slot in our minds over evolutionary time. What defines this entry – and what is common to all of these relationships – is that agents will occupy these roles with respect to one another.

The order of causation is on-the-ground relationships and cost/benefit calculations first, summary representation second. Individuals are making cost/benefit decisions over evolutionary time, based on a number of factors, such as self-interest, caring about the welfare of another, a formal social contract, and so on. These first-person decisions then create an invariance in the world – a class of events (in this case it is the occupation of group-constitutive roles) – that can in turn select for a conceptual entry that summarizes this class (Shepard, 1994).

This summary representation allows agents to communicate about and refer to this class, both within themselves, and also with others. For instance, an agent can observe their own behavior and motivations, and create a summary representation of who they are in a "group" with, and who they are not. Or they can observe other's behaviors and do the same. Of course, by "agent" we mean a set of information-processing functions (as opposed to a conscious agency within the mind; see Pietraszewski & Wertz, 2021). As such, there is no entailment that each (token) summary representation becomes conscious and communicable to others, although undoubtedly a subset does.[5] These communicable summary representations (which would in principle be highly compressed signifiers) should, when communicated, induce an expectation of the class of events to which they refer. That is, if one is told that X and Y are in a "group" within a conflict, then this should lead to the expectation that X and Y will be more likely to occupy the above roles with respect to one another, all else equal.

5. An example: Understanding group-constitutive roles from the perspective of a particular agent

What does it mean that group-constitutive roles *are* group membership to the human mind? In the examples that follow, we will consider how group-constitutive roles play out from the perspective of a particular agent. These observations, which are entailments of our computational theory, describe additional elements of the overarching information-processing problem of representing, reasoning about, and carrying out the group-constitutive roles. For instance, we will see that occupying group-constitutive roles requires our focal agent to sometimes be motivated to act, to anticipate the actions of others, and to make certain cost/benefit calculations.

We will begin with Figure 3. In the top panel of Figure 3, a third agent, denoted C, observes A attack B. If this agent is to become involved in the conflict, then some constellation of the

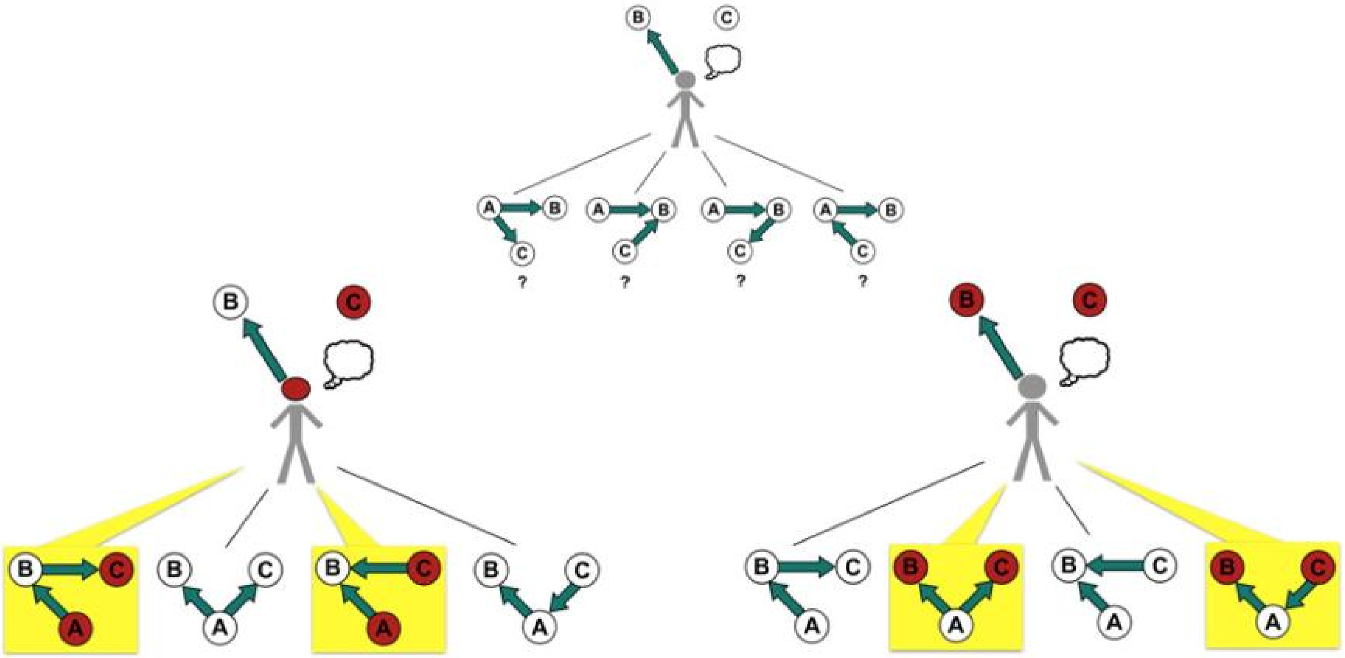


Figure 4. The single agent is now in role A – the cost imposer. Again, without knowing group membership we have no basis for anticipating which, if any, of the four possible triadic interactions are likely to happen (top panel). However, once group membership is marked (in red; bottom panels), certain next steps become more likely than others (those highlighted in yellow).

four possible triadic interaction types will occur. However, as outside observers who do not know if they share a group membership with either A or B, we have no basis for narrowing down which will occur. In contrast, once we know that the onlooking agent is in a group with A and not with B (Fig. 3, bottom left panel), only two of the four interaction types become likely, and the other two become less likely. In contrast, if the onlooking agent is in a group with B and not with A (Fig. 3, bottom right panel), then the other two triadic interaction types become likely. Thus, the representation of the background relationship – the "group" membership – between the direct interactants (A and B) and the onlooker (C) modifies which behaviors are most likely to follow from A attacking B.[6]

Notice that for these triadic interactions to occur, the initially uninvolved onlooker must at times behave. For example, if the uninvolved agent C shares a group membership with A, then they will join with A in attacking B (the first highlighted interaction type in the bottom left panel of Fig. 3). This means that the psychological implementation of these group-constitutive roles requires motivations to act: In this case, group membership is a relationship between A and C that is causing C to be more likely to act against B. Notice, too, that C must sometimes also generate expectations of other's actions. For example, C in this example is also more likely to be attacked by B, given that A and C are group members (the second highlighted interaction type in the bottom left panel of Fig. 3). This means that the psychological implementation of these group-constitutive roles also requires producing expectations about others' behaviors.

The interplay between one's own behavior and the expectations of other's behavior can also be seen in Figure 4, which again depicts group membership from the perspective of a single individual, who is now in the role of A, the attacker. What happens after A attacks B will depend on the group membership of the uninvolved onlooker C. If that onlooker is in a group with the attacker, then their presence is a benefit from the attacker's perspective (bottom left panel of Fig. 4); they may also join in attacking the victim, helping diffuse the costs and risks of the conflict. Or the victim may retaliate against the attacker's fellow group member rather than the attacker themselves, thereby diffusing the risk of counterattack. In contrast, if the onlooker is in a group with the victim, the likely next steps are costly to the aggressor (bottom right panel of Fig. 4): The aggressor will either be motivated to also attack the onlooker, requiring additional cost and risk. Or, in response to the initial attack on their fellow group member, the onlooker will in turn attack the aggressor. Neither event benefits the aggressor. Consequently, if aggressor A can anticipate these outcomes *before* attacking B, then A will be more likely to attack when A's group member is present (right panel), but will be less likely to attack when B's group member is present (left panel). This is the causal logic for why the presence or absence of group members changes decision thresholds for engaging in or avoiding conflict. This example demonstrates the strategic benefits of representing the group membership of uninvolved third parties, and how one's own decision-making is aided by being able to anticipate how others will behave.

Finally, Figure 5 depicts the perspective of the initial victim. Here, the initially uninvolved onlooker can either be a group member with B (left panel) or with A (right panel). Again, there is a cost/benefit asymmetry between these two scenarios: It will be less costly for the victim to be attacked in the presence of a fellow group member (left panel) than in the presence of a group member of the attacker (right panel). Furthermore, our victim B will sometimes be motivated to act (e.g., attacking C when A and C are group members) and will sometimes anticipate other's actions (e.g., expecting C to attack A on his or her behalf when B and C are group members). For detailed cost/benefit analysis of why these particular events may unfold, see Pietraszewski (2016, SOM).

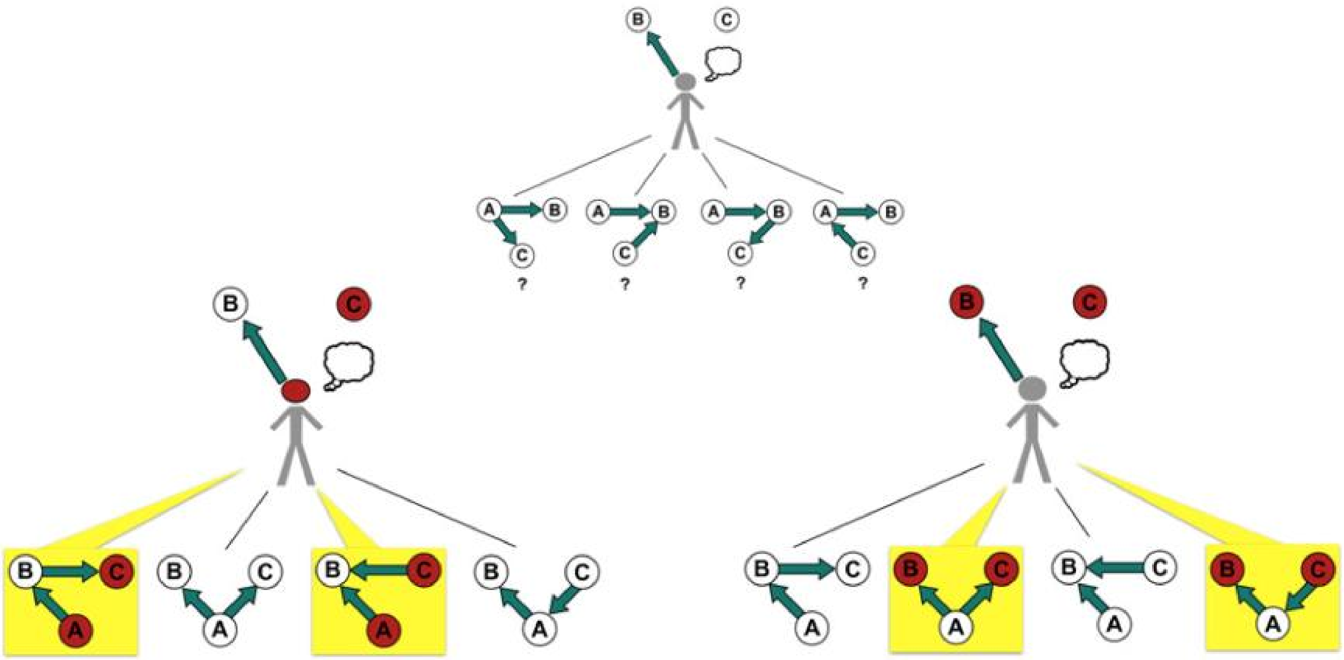


Figure 5. The single agent is now in role B – the victim of the initial cost imposition. Again, with knowing group membership, we have no basis for anticipating which, if any, of the four possible triadic interactions are likely to happen (top panel). However, once group membership is marked (in red; bottom panels), certain next steps become more likely than others (those highlighted in yellow).

6. Clarifying the definition

We can now refine and clarify the definition: Roughly, somewhere in the mind are floating around the representations [group] and [conflict]. When these co-occur, what is bound to this composite [group-in-the-domain-of-conflict] representation are these group-constitutive roles within the triadic primitives. That is, these roles are the mind's operational or functional definition of a group in the context of conflict, and these triads are the skeletal primitives out of which each instance (*token*) of a group is perceived and built (and we will see in a moment how all of this scales up). In other words – and tautologically, given our definition – to the degree agents belong to the same group, they will be more likely to act according to these roles, and will also anticipate that other group members will do the same.

This definition entails a minimal bound for what the mind considers a group (conflict) dynamic: Minimally, two agents must be within the group, and a single agent must lie outside of it (where "group" is defined non-tautologically by the occupation of group-constitutive roles). For example, if we imagine a conflictual world with only three agents, in which two are allied and a third is not, then we would see over time the two allied individuals occupying each of the roles that define a group when engaging in conflict with the third lone individual. The two in conflict with the lone individual would constitute a group, according to the definition.

Furthermore, the group-constitutive roles define "group" from a third-party perspective. Agents who are not involved in a particular conflict can nevertheless infer these roles by observing the actions of others, and who occupies these roles will define for these third parties who is in a group with whom. Of course, the roles are mental representations and so cannot just be seen (in the same way that an object cannot just be seen; Marr, 1982). Therefore, a substantial amount of input analysis is required to go from the detection of on-the-ground behaviors (such as seeing A shooting at B, hearing about X gossiping about Y, and so on) to the inference that these behaviors are an instance of one of these roles; a problem that is far from trivial. Moreover, contingent cost-infliction is often a drawn-out process, with many gaps and lulls between interactions. For example, if you steal my cattle today, it may be a day, a month, or even a year before I retaliate. Therefore, delays between cost-inflictions will have to be allowed for, as will some opacity of who has done what to whom. It follows then that identifying who is in what group requires understanding who has imposed which costs on whom, and in response to what past behaviors. Inferring group membership is then in part an attribution problem.

Crucially, behaviors will not be the only cues used to infer the occupation of the group-constitutive roles. In particular, because the relationships that cause agents to occupy group-constitutive roles with one another constitute an evolutionary invariance (at a particular level of abstraction), natural selection can "see" what cues correlate with these relationships, and can build learning systems that spawn probabilistic group-constitutive role representations in response to them. Consequently, even very sparse cues that do not have much to do with conflict when looked at superficially (e.g., Dunham, 2018; Pietraszewski, 2013, 2020) can nevertheless probabilistically elicit these group-constitutive role representations. We will refer to these cues as *ancillary attributes* of groups (ancillary because they are not behaviors corresponding to the group-constitutive roles themselves, but rather are cues that predict the occupation of these roles, either over evolutionary or developmental time).

What attention to ancillary attributes means, is that even in a world in which the *only* reason to attend to "groups" is to predict conflict-related outcomes, group tokens that superficially have little or nothing to do with conflict will nevertheless be attended to: a point that has interesting consequences for our understanding of the folk concept "group" – namely, it raises the issue of whether or not the concept "group" can ever be divorced from conflict-related representations no matter how apparently peaceful the group token is. In practice, both ancillary attributes and behaviors corresponding to group-constitutive roles are likely to be present in the world. An additional information-processing problem is to then link up these diverse representations to the correct group token.

In sum, the information-processing within the mind that makes it possible to consider, plan, and generate the above group-constitutive roles is what constitutes the psychology of groups-in-conflict. Furthermore, this psychology is recursive in the sense that I not only expect group members to occupy these roles, I expect that others will also expect this, and so on. Finally, we have so far been treating behavior as all-or-none – that if someone is a group member with someone else, they will produce behaviors X, Y, and Z. Of course, this is an oversimplification, so more fleshed out computational theories can incorporate something like a continuously scaled threshold to act, as part of a larger "motivational" goal-tree/difference-reduction architecture.

7. Scaling up

The architecture for representing each of the triadic primitives is based upon interactions between individual agents. However, at times, multiple agents will all behave the same way within a particular triadic primitive (e.g., a set of agents will all occupy role C in *Defense*). The architecture can therefore scale up the representation by inserting (or substituting) that entire set of agents into the single-agent slot (the group-constitutive role). This one-to-many substitution allows the same (or a slightly modified) architecture to handle much larger numbers of agents simultaneously. Moreover, to the degree that that particular set of agents are in a group – per the definition – they are permitted to be inserted into *all* of the agent slots within all of the triadic primitives depicted in Figure 2, which enables an infinitely hierarchical scaling up of nested group representations. This kind of architecture allows us to represent, for example, that Japan and Germany occupied group-constitutive roles with respect to one another during World War II, as did Britain and the United States, without having to calculate the actions of the millions of individuals involved (see also Fiske & Neuberg, 1990; Insko & Schopler, 1987).

Of course, any one of the roles within these triadic primitives may also be produced for reasons other than group membership. For example, an agent, C, who witnesses A attacking B, may retaliate against A, not because she is a group member with B, but because A's attack was morally reprehensible, and should be punished. Indeed, even if multiple agents all decided to attack B, thus filling the role C in *Defense*, this would still not constitute a group. Groups, instead, are a special class of relationships in which *all* of the above roles either will occur, or are expected to occur. That is, group membership applies to all of the triadic primitives. If only a subset are shown to hold, this should hurt the representation that such a relationship is a group.

For example, if we were to demonstrate that all of the agents filling role C in the above example are not willing to place themselves in any of the other group-constitutive roles with respect to A, aside from this one case, then they would not be seen as constituting a group with A. That is, they are not group members precisely *because* these agents do not occupy all of the above roles with respect to one another. Indeed, different collections of agents can occupy some, all, or none of the above group-constitutive roles with respect to one another. This variation must be represented by the mind, and this represented variation may explain why there is a perceived continuum of "groupishness" or entitativity across different collections of people (Campbell, 1958; Hamilton et al., 2002; Ip et al., 2006; Pietraszewski, 2016) – an idea that is imminently testable.

8. The utility of the definition

So what does having a plausible theory of a group representation in the context of conflict buy us?

8.1. Ancillary attributes

First, the present account suggests that, for a better part of a century, psychologists have been studying groups largely through the lens of ancillary attributes that groups tend to have, rather than by directly studying the fundamental behaviors (or more precisely, the intergenerationally recurrent dynamics) that constitute group membership itself. *Ancillary attributes*, to review, are those things that help diagnose the occupation of group-constitutive roles: they are cues that predict or correlate with these roles (either over evolutionary or developmental time), but are not the on-the-ground cost-imposing behaviors that correspond to the group-constitutive roles themselves.

Allport, for example, defined a group as "two or more persons who are assembled to perform some task, to deliberate upon some proposal or topic of interest, or to share some affective experience of common appeal" (1924, p. 260). The problem with this definition is that groups can exist without assembly, deliberation, or a shared affective experience. Now of course it may be true that these things do tend to co-occur with the formation and maintenance of groups (precisely: the classes of relationships that we are referring to as groups), and our psychology should take advantage of this fact – but they are not necessary. In fact, it is trivial to think of counter-examples to such ancillary-attribute-based definitions: A collection of students who are all assembled to take a test are less of a "group" than are a set of students all assembled together to burn the books of the locally scapegoated outgroup, for example. The latter behaviors diagnose a disposition toward behaving toward that outgroup in a way consistent with our definition, whereas the former does not. Other examples of ancillary attributes include spatial proximity, overall similarity along some dimension, and so on.[7]

Historically, psychological theorizing on what defines or constitutes group membership has been largely focused on these ancillary attributes (e.g., Campbell, 1958) – undoubtedly because they are visible. Much of the same can be said for empirical research (e.g., Pietraszewski, Curry, Peterson, Cosmides, & Tooby, 2015; Powell & Spelke, 2013). This is not to say that these ancillary attributes aren't important. They are. Instead, they are simply not always necessary nor sufficient, meaning that in a deep sense that they diagnose group membership, but they are not what group membership is fundamentally about. Notably, if one were to observe a collection of people who share the same arbitrary marker or "affective experience," for example, but who do not occupy group-constitutive roles within a conflict, then that marker and that experience will not be seen as constituting group membership. In contrast, the present definition is sufficient on its own: Even if a collection of people have no ancillary attributes – for example, are not physically proximate and have no other similarity or shared emotional experience – they will nevertheless be viewed as a group, so long as they are shown to occupy each of the roles in our definition.

This focus on ancillary attributes is no accident. Rather, it is a consequence of a scientific history of relying on intuition:

Intuition highlights what is *variable* about groups (who belongs to what group, and what individual tokens of groups exist) while blinding us to what is *universal* about group memberships (what constitutes a group, and what is done within cognitive architecture once a group is detected). This is why group psychology has been largely treated as a categorization process – as if the assignment of agents to groups were the main issue – whereas in fact conceptualizing *what* exactly agents are being assigned *to* is the far deeper and more fundamental issue.

8.2. Computational adequacy – moving beyond containment metaphors

Second, having an explicit computational theory forces us to address issues of computational adequacy – something missing in the group literature until now. (*Computational adequacy*, to review, is concerned with asking if the information-processing proposed is sufficient to account for the observed characteristics of a particular phenomenon.)

One consequence of this fact is that nearly all theorizing about groups has relied (either explicitly or implicitly) on a containment or subsumption metaphor – that groups are containers into which individual members are placed (for the reasons just described above). In contrast, the present account suggests that group membership is a *relational* property (specifically, the relative strength of pairwise comparisons among agents; DeDeo, Krakauer, & Flack, 2010; Perry, Barrett, & Manson, 2004). On this account, who “belongs” to what group is borne out of a calculation of the relative relationships among the agents involved (which is why – to put it pre-theoretically – who one feels closest to depends on who else is around).

Moving beyond the containment metaphor is important because the metaphor fails to account for even the most basic characteristics of social groupings – such as that when multiple, nested groups exist, switching between group identities can and frequently does occur. For example, two populations may be on opposing sides during a civil war (i.e., each side will occupy *non*-group-constitutive roles with respect to one another). But if both populations are then attacked by a larger common enemy, the two populations may subsequently occupy group-constitutive roles with respect to each other *against* that common enemy – meaning that the two populations that had not been in the same group now are.

Although an obvious example, containment metaphors cannot account for why group membership is a function of who is interacting: When taken literally as information-processing accounts, they simply produce assignments of each agent to each category. There is no mechanism to switch between the containers, nor to inhibit one in a favor of another (see also Liska, 1962; Von Neumann & Morgenstern, 1944). In contrast, the triadic primitive architecture here performs relational pairwise comparisons among triads of agents, calculating which two are more similar *among the agents represented within the triad* (which is the crucial step that allows agents on opposing sides in the civil war to be represented as members of the same group for the purposes of interacting with the larger common enemy – a calculation strictly speaking not possible with the containment metaphor[8]).

In contrast, the present definition anticipates these shifts and extracts meaning from them: Which group membership (or identity) is “relevant” or “active” is that which is currently aligned with the group-constitutive roles being currently expressed. Each group representation necessarily carries with it a set of behavioral expectations directed toward specific others, and each individual can be a member of any number of different groups. Therefore, *which* group membership is relevant at any one time is a function of with whom the agent is interacting.

Furthermore, agents may not be able to resolve all of their group memberships – in the sense that some group representations can be activated simultaneously, producing incompatibilities about which group membership should be the basis of one’s current behavior. The metaphorical language of feeling “torn” about opposing group loyalties occurs in such cases because each group membership implies a specific class of behaviors directed at particular others. For example, an agent in role C, observing A attack B, may be a member of one group that would require attacking B, and another that would require attacking A. “Feeling torn” then corresponds to when an agent belongs to at least two groups that would, according to our definition, require acting in different ways toward those others. Group-constitutive roles specify what representations agents who have conflicting group loyalties are wrestling with, and must make a decision about. The definition thereby allows us to conceptualize such felt conflicts as a class of information-processing outcomes.

8.3 Integrating group and individual decision-making, and accounting for what social identities are and why they matter

Third, by defining group membership as changes to internally represented thresholds for producing and/or expecting cost impositions, our definition allows us to integrate the concept of “group membership” with an individual’s other cost/benefit calculations and decision-making. This has not been possible before. For example, theorizing that relies on the metaphorical language of individuals becoming “bonded” or “fused” to a group evokes the right idea of an agent acting on behalf of a particular identity, but cannot be quite literally correct, because the agent and the group never become exactly the same entity.[9] Even young children understand that allies are not voodoo dolls, in that they do not simultaneously experience all of the same internal states simply because they are bonded or associated together (Pietraszewski & German, 2013). Adults, furthermore, understand that one’s social alliances can be in conflict with one’s own independent cost/benefit evaluation of a particular situation (Pietraszewski & German, 2013; Shaw, DeScioli, Barakzai, & Kurzban, 2017). Certain theories argue that morality itself exists specifically for this category of eventualities – freeing agents from their own alliance obligations and allowing them to impartially evaluate events from the perspective of disinterested third parties (DeScioli & Kurzban, 2009, 2013).

Until we can conceptualize group membership as a set of continuous decisions on the part of individuals, it is not possible to study how conflicting interests within an individual, or between individuals within a group, are integrated and resolved within the mind (a rather serious deficiency – as this is one of the central elements of what we want to understand when we study group conflict). Past accounts that treat groups with a containment metaphor do not have a literal model of how individual decisions and groups relate to one another (nor how different group memberships are resolved – e.g., how one chooses between tribal and religious loyalties). On these past accounts, it is of course possible to *note* that conflicts of interest and disloyalty exist, but such phenomena do not fall out of the information-processing logic of what a group consists of in the first place, and as such fail to be accounted for. In contrast, if we understand “group” to emerge out of individual decisions within sets of triadic interactions, these conflicts of interest and a number of other phenomena fall out naturally (see Pietraszewski, 2016).

A similar issue arises with respect to the phenomenon of *social identity*, which is arguably one of the most important contributions to come out of social psychological research in the twentieth century. Briefly, early theorizing about intergroup conflict assumed that conflict would be based, roughly, around conflicts of interest. This was called *realistic conflict theory* (e.g., Levine & Campbell, 1972). What social identity theorists discovered was that objective reasons for conflict were not necessary to spawn "groupishness," even in the lab (e.g., Sherif et al., 1961; Tajfel, 1970, 1982; Tajfel & Turner, 1979, 1986; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987). Instead, people seemed to intrinsically value social relationships, and were willing to pay a cost to "have" them (i.e., they would coordinate their actions according to abstract identities, rather than to objective aspects of the pay-out structure of the world; for analysis and reviews of this family of theories and findings, see Bar-Tal, 2001; Brewer, 2001; Dawes, Van De Kragt, & Orbell, 1988; Dunham, 2018; Park & Van Leeuwen, 2015; Pietraszewski, 2020; Tajfel, 1982).

An adequate account of the psychology of groups must account for social identities (meaning that agents can have them, and can represent them in others). Yet it has been rather difficult to integrate social identities in literal, mechanistic models of on-the-ground social interaction (for valiant attempts, see Smaldino, Pickett, Sherman, & Schank, 2012). For the most part, models have either assumed away social identities or have added them ad hoc, such that their importance does not emerge from the structure of the individual interactions within the model, but is declared by fiat.

In contrast, the present proposal articulates what causal role social identities play within the mind's information-processing: they are placeholders that cause agents to treat each other as substitutable with respect to one another within the triadic primitives. That is, a social identity is a mental representation (real or imagined) that leads to the scaling up from a single agent to a set of agents within one or more of the agent slots within the triadic primitives. For instance, if a person believes that an attack was caused by a single individual, the contingent response to that attack will only be directed at that individual. However, if a person believes that an attack was caused by a social identity, then that person may retaliate against others also holding that same social identity, even if they were not involved in the initial attack (because from their perspective, the two agents holding the same social identity are substitutable with one another).

This framework allows us to understand why social identities are of value, and why people are willing to work so hard to maintain and change other's representations of them (e.g., Haslam, Reicher, & Platow, 2011): Representations of who and who does not belong to the same social identity become powerful determinants of how particular conflicts spread, and who retaliates against whom. It is this down-stream effect of extending a conflict out onto many more previously uninvolved agents that explains why a conflict involving social identities activates additional attribution processes (Brewer, 2001; Dawes et al., 1988) and produces more fraught and protracted conflict dynamics than those that do not (Abrams & Rutland, 2008; Bar-Tal, 2001; Puurtinen, Heap, & Mappes, 2015; Walter, 2004, 2009).

Moreover, individuals can hold different perceptions of what social identities exist, even when directly interacting with one another. For example, history is replete with bloodied colonial powers assuming away the heterogeneity of local, native social identities and treating everyone as if they are the enemy (such as retaliating against villagers for raids by rebel groups in the forest), which eventually becomes a self-fulfilling prophecy. The current framework allows us to understand what it means, precisely, that different agents hold these diverging representations, and what work, in principle, these representations are doing in each of their minds.

8.4 Accounting for group emergence

Fourth, a literal notion of what it means for each agent to represent themselves and others as belonging to the same group within a conflict offers a principled way to think about group emergence. For example, the current proposal suggests that groups (or more precisely: increases in the degree to which agents are probabilistically assigned to group-constitutive roles) can emerge out of individual actions, such that their existence may be unplanned until some event transpires.

For instance, suppose that we are one of a collection of people waiting around for a bus. This collection is not yet, on any meaningful definition, a "group." But suddenly someone drives by and throws a stone at one of us. We can either (i) throw a stone back at the car (now intuitively "more" of a group, as we are occupying role C in *Defense*), (ii) hit the victim (a different group, as we are occupying role C in *Alliance*), or (iii) do nothing (non-group).

Before we act, each one of these possibilities is equally possible and realizable. Yet it is only when we act (or, perhaps, make a decision to act) that certain groupings now become more likely going forward. Moreover, we must also coordinate – a single decision (i.e., a single triadic interaction type) does not make a group, particularly if no one else agrees with our assessment. For example, we may throw a stone at the car, believing that we are now in a group with the victim. Whereas the victim turns and hits us, contradicting our assessment. Certain behaviors can therefore be seen as bids for establishing group memberships, precisely because group memberships are in a deep sense a set of behaviors.[10]

While this stone-throwing example is oversimplified, something roughly analogous to this de-centralized dynamic often plays out in the genesis of real-world groupings. The current framework, therefore, suggests a number of research programs with the goal of viewing bids at group formation – such as political rhetoric and narratives of historical grievances (e.g., Lopez, 2020; Moncrieff & Lienard, 2019) – through the lens of the triadic primitives and group-constitutive roles.

9. Next steps

9.1 An engineering approach to the study of groups, and a more in-depth task analysis

We can now return to our hypothetical from the beginning of the paper, in which an AI engineer approaches us and wants to build a machine intelligence capable of "groups." What exactly would we tell them? Our computational theory of a group in the context of conflict now provides them with a clearer, more literal set of information-processing functions to be implemented in mechanisms. These include

- Machinery for stringing together long chains of concatenated triadic primitives, in which the same agents are able to be assigned to different agent slots within each triadic primitive (Bob may be in role C in one, A in another). These chains can be thought of as a path through a much larger *triadic state space* (e.g., see Pietraszewski, 2016). These chains describe the most likely interactions that will occur among agents as a conflict unfolds, and are potentiated by (and thus are tagged to) *agent-based frames* and *event-based frames* (the former being concerned with assessing what will happen if particular agents interact, whereas the latter is concerned with assessing what will happen if a particular event occurs.)

- Machinery for storing and making use of "group" representations, which (as shown in the examples of Figs. 3–5) collapse large portions of this triadic state space, modifying which paths are most likely. That is, group representations are not containers, they are representations (namely, modifiers) that modify what would otherwise be the defaults of this triadic state space, potentiating certain areas of this space, and curtailing others.
- Machinery for updating group representations and most-likely triadic-state-space representations, based on ongoing events in the world (e.g., the actions that I take and the actions that others take).
- Machinery for generating agent-based and event-based counterfactuals (e.g., what would happen if I took action X, what would happen if Bob took action X, what would happen if I run into Bob, what would happen if Bob runs into Bill, and so on), the outputs of which should be stored in a manner to allow for easy/random access, and would in principle feed into motivational and planning systems.

The above are just a handful of the minimal information-processing requirements for something like a set of mechanisms capable of using group representations as defined in the current proposal. There must also be mechanisms for scaling up and scaling down based on social identities, for quarantining counterfactuals from actual events, for translating cues in the environment into probabilistic group-constitutive role representations (and for translating these representations into concrete behavioral predictions), for identifying disloyalty and distinguishing between disloyalty toward a group and the non-existence of a group, and so on. This is exactly what we want: The point of having a computational theory is that it reveals information-processing problems to be solved.

The information-processing requirements (or problems) outlined above represent a core element of humans' everyday commonsense knowledge – an issue of central importance within current AI (e.g., Etzioni, 2018; Mueller, 2015). For example, it is obvious why two complete strangers who have never met – someone wearing an axis uniform who was born in Marburg and someone wearing an allied uniform who was born in Iowa city – are shooting at each other when we look at grainy footage from World War II. Just as it is blindly obvious that someone is taking their safety in their hands by cheering for Manchester United at Liverpool or for the Yankees at Fenway. Yet no robot or AI currently holds such intuitions. This is because these intuitions are the outputs of the above-listed information-processing steps, which can be thought of as a series of social inference engines (akin to a physics engine; Ullman, Spelke, Battaglia, & Tenenbaum, 2017), without which an appreciable swath of commonsense knowledge about the social world would not be possible.

The above-listed information-processing steps are also an important part of the solution to the *frame problem* within the social domain (i.e., for representing what will change in the world when action X is performed; Kamermans & Schmits, 2004; for an initial empirical exploration, see Pietraszewski & German, 2013) – and something like these inference engines will need to be implemented in some kind of event calculus (Mueller, 2015) or any other kindred architecture that permits deduction (i.e., prediction and explanation) while avoiding monotonicity (i.e., "combinatorial explosion": Minsky, 1974). The event types presented here (a verbal and graphical description involving contingent cost imposition) are but a sliver of a fully computationally adequate account, and so our description of the event calculus itself is thus far highly cursory. Nevertheless, the goal here is to provoke the fleshing out everything that would be needed (i.e., a task analysis).

One of the primary goals of this paper, then, is to invite the broadest possible community of researchers – particularly those who work in AI, software engineering, computer and cognitive science, and so on – into the study of social groups, and allow them to ask (in collaboration with researchers already studying social groups): What are all the things that one would need to put into an information-processing entity, such that it could go out into the real world, observe and interact with that world, and then generate and use the particular representational system presented in this current proposal? There are likely hundreds, if not thousands of information-processing functions and subfunctions required to make all of this happen. Enumerating what these functions are will provide us with a *computational ontology* – a set of concepts that describe information-processing entities and their relationships with one another (e.g., Mueller, 2015; Russell & Norvig, 1995), and will define what we should be looking for within the mind (or implementing in automata). A complete understanding of the psychology of social groups will require proposing, testing, and then either confirming, refining, or correcting every single element of this ontology (along with many others) – an explicitly computational-theoretic approach that will constitute a mutually informative *Cognitive Psychology* and *Cognitive Science* of social groups. Despite the incredible social importance of understanding the psychology of group-based conflict, it is not clear that any such research program is currently underway.

9.2 Empirical predictions, methods, and outstanding questions

On the experimental side, our computational theory makes a number of empirical predictions that await testing. For example, the behaviors that correspond to what we have called *group-constitutive roles* should be inputs that lead to the perception of groups-in-conflict. That is, if one were to show instances of cost-infliction behaviors among triads of agents, then participants should perceive those agents occupying group-constitutive roles as being members of the same group. Likewise, group-constitutive roles should be also outputs, in that they should be expected patterns of behavior once a summary representation is communicated. For example, if we tell participants that Harry, Ron, and Hermione are all in a group, and that they are in conflict with Malfoy, Crabbe, and Goyle, who are all also in a group, then participants should be able to predict which roles these agents will occupy with respect to one another. In other words, the logic of the definition should be intuitive to human participants, just as Figures 3–5 should have been intuitive to the reader. Furthermore, the use of these group-constitutive roles as both inputs and outputs should be universal and hold for all humans on earth.

The relationship between more direct cues (i.e., behaviors that clearly denote group-constitutive roles, or verbal labels referring to a known summary representation) and what we are calling ancillary attributes of group membership will also be important to establish. That is, ancillary attributes or cues (i.e., things that are concomitants of the occupation of the group-constitutive roles, such as similarity, proximity, etc.) should be used to probabilistically infer group membership in the absence of more direct cues. A number of as-yet unanswered questions then arise. For example, how do direct cues fare against ancillary cues? In principle, ancillary cues should be used when no other direct cues are present, but should be secondary to direct cues.

Another question is what happens when ancillary cues are pitted against direct cues. The current proposal predicts that direct cues should corrode the validity of ancillary cues. How, in turn, are ancillary cues learned, and do some occur independently of learned associations? For instance, some attributes such as proximity and shared opinions may be intrinsically linked to guesses about group membership (e.g., Gershman & Cikara, 2020; Gershman, Pouncy, & Gweon, 2017; Lau, Gershman, & Cikara, 2020; Pietraszewski, 2013), whereas other attributes – such as how one ties one's shoe laces, which is sometimes used by gang members – may depend much more on social context. The proposal here would predict that arbitrary similarities should become interesting to the mind and be seen as "group" markers insofar as they track or correspond to group-constitutive roles (see also sect. 6, "Clarifying the definition," above). Indeed, there is already some experimental evidence for some of these predictions floating around in the past literature (e.g., Bar-Tal, 2001; Cikara, Bruneau, Van Bavel, & Saxe, 2014; Cikara, Van Bavel, Ingbretsen, & Lau, 2017; Ip et al., 2006; Lau, Pouncy, Gershman, & Cikara, 2018; Patton, 1996, 2000; Pietraszewski, 2013; Pietraszewski et al., 2015; Rhodes & Chalik, 2013), but the full weight and entirety of these predictions remain to be tested.

Finally, because the present proposal is explicit about what the end-state representation in the mind needs to be (that a "group" token out in the world become yoked to either direct or ancillary cues, which are in turn yoked to group-constitutive roles), a principled *learnability analysis* of the social environment can be conducted. That is, one can understand when and why particular tokens need to be apprehended via ancillary cues (e.g., because the direct cues are infrequent, dangerous to observe, and so on).

There are also a number of unresolved empirical issues brought up by the present proposal. For instance, a cost/benefit analysis of each of the group-constitutive roles within the triadic primitives suggests that some roles will be more costly than others. For example, defending a victim against an aggressor is less costly than helping an aggressor attack a victim, all else equal (for the explanation, see Pietraszewski, 2016). This means that *within* the group-constitutive roles, some may be more likely than others. Therefore, are some more anticipated than others? And are those that are more costly the least expected? Asymmetries between offense and defense, for example, have been found (Böhm, Rusch, & Gürerk, 2016; De Dreu et al., 2016; De Dreu & Gross, 2019; Lopez, 2017; Rusch, 2013, 2014a, 2014b), and the present proposal suggests a number of new comparisons to test in future studies. Likewise, the relationship between the different group-constitutive roles should also be probabilistic – meaning that seeing two agents occupying group-constitutive roles in one triadic primitive may increase the expected probability that they will also occupy group-constitutive roles within another triadic primitive (see also sect. 8.4, "Accounting for group emergence"). The mind, in other words, is expected to guess about group membership based on partial information – an idea that is imminently testable.

10. Toward a complete computational theory of social groups – including non-conflict-based representations

Finally, because agents can also coordinate and cooperate with one another – and there is nothing in our conflict-based event types to allow for the representation of such events – the current theory of what constitutes a group representation is not yet complete. Therefore, even if correct, the present account will need to be supplemented with (at minimum) additional event types – and there are a number of existing theories and taxonomies that can be repurposed as plausible hints at what these types might be (e.g., Alexander, 1987; Balliet, Tyber, & Van Lange, 2017[11]; Heider, 1958; Tatone, Geraci, & Csibra, 2015).

Although the current proposal is silent on what these might be, we can at least speculate that a plausible (if not obvious) set of candidate event types for something like polyadic cooperation might be something akin to sets of *direct reciprocity* (A gives to B, B repays in kind over time) and *indirect reciprocity* (A gives to B, as a consequence, B gives to C; see Alexander, 1987). However, this is just speculation, and there may need to be additional elements to make these computationally adequate, or something more specific may be necessary – issues that await future work. What we can say with some certainty is that neither direct nor indirect reciprocity can account for the current conflict-based event types (i.e., these are not isomorphic with respect to one another)[12] – which means that a complete computational theory of groups will likely be comprised of multiple event-type frameworks.

In the search for a complete computational theory of group representation, care will have to be taken to conceptuallydistinguish between two different enterprises: The first is to provide a computationally adequate account of every group token. In this respect, each token will likely be a composite of a number of event types (in the same way that an object representation is a composite of visual information-processing types, like lines and colors). A second enterprise is to explain why people can think about and understand the over-arching category "group," and why they agree about a continuum of groupishness across these tokens. The first enterprise will require a proliferation of types. The second will require some unification or integration of those types. There are many more *reduction problems* to be solved, in other words, than just the problem of multi-agent conflict dynamics covered here. The principle of computational adequacy – combined with considerations of what is both necessary and sufficient – will have to guide these future efforts

11. Summary and conclusion

Despite an enormous literature on groups and group dynamics, little attention has been paid to explicit computational theories of how the mind represents and reasons about groups. The goal of this paper has been, in a conceptual, non-technical manner, to propose a simple but non-trivial framework for starting to ask questions about the nature of the underlying representations that make the phenomenon of social groups possible – all described at the level of information-processing. This computational theory, when combined with many more such theories – and followed by extensive task analyses and empirical investigations – will eventually contribute to a full accounting of the information-processing required to represent, reason about, and act in accordance with group representations.

Acknowledgments. The author would like to thank Mina Cikara, Carsten De Dreu, Florian van Leeuwen, William von Hippel, and several anonymous reviewers for helpful comments on previous versions of this manuscript.

Financial support. The author was supported by the Max Planck Society while preparing this manuscript.

Conflict of interest. None.

Notes

1. In most cases, researchers understand the assume-away nature of these games, and draw appropriate inferences from their results. These describe what in addition to group representations may influence group dynamics (e.g., reputation heuristics; Gross & De Dreu, 2019). Also, for a notable contribution toward a computationally adequate model within this literature, see DeDeo et al. (2010).
2. *Computational theories* are not the same thing as *computational models* (Palminteri, Wyart, & Koechlin, 2017). *Computational models*, which are tools for modeling complex phenomena, can be applied to anything – including the modeling of entities that do not process information as part of their function, such as weather patterns, stresses on bridges and buildings, and so on. *Computational theories*, in contrast, describe what is going on within some information-processing entity (such as a computer, cash register, or mind) when operating (Marr, 1982). That is, they are "theories of the information-processing" that underlies a particular phenomenon (e.g., Chomsky, 1980; Marr, 1982; Minsky, 1961, 1974; Tooby & Cosmides, 1992). Although computational theories can be made into computational models (e.g., Palminteri et al., 2017) not all computational models, even of the mind, are *computational theories* in the sense of approaching the mind at the level of information-processing (see also the "ad-hoc, heuristic" approach described in Marr, 1982, p. 19).
3. In philosophical parlance, we are thereby providing a *functional* (or conceptual or procedural) *role semantics* – in which a linguistic token or folk-psychological construct corresponds to a state within an organism, and that state is ultimately defined by its observable, functional effects (Block, 1997; Dennett, 1969). That is, we are here defining a group representation with respect to its functional role within the rest of the cognitive architecture.
4. "Evolutionary time" refers to time spanning multiple generations (i.e., phylogenetic time), as opposed to time spanning a single lifetime (i.e., ontogenetic or developmental time). The crucial difference being that evolutionary/ phylogenetic time is the time-course over which natural selection (differential reproductive success) occurs, thereby shaping the structure of information-processing adaptations, whereas ontogenetic time is the time-course over which each token of an information-processing adaptation is expressed developmentally and used (e.g., Barrett, 2015; Scott-Phillips, Dickins, & West, 2011; Tooby & Cosmides, 1992).
5. One might reasonably wonder why the existence of such summary representations would be necessary, aside from the fact that the mental concept "group" is the thing-to-be-explained (the explanandum). That is, why does the mental concept of a group exist at all? One answer is that the summary representation permits communication, allowing us to think and communicate divorced of the on-the-ground cues that would otherwise have to be seen first-hand. For example, if I tell you that Sue and Steve belong to group X, then you can know what that means without having directly seen or inferred the on-the-ground cues that caused me to think they were in a group in the first place. The summary representation thereby allows us to communicate generalities that can be applied to multiple, more specific situations, without having to be tied to all of the specifics. A second, more fundamental answer to why summary representations are likely necessary has to do with the nature of computation itself: that communication among representations *within* the mind is just as important as communication between individuals outside of it. Summary representations provide content-searchable tags that help solve the *integration problem* of delivering the right information to right place at the right time (Lorenz, 1948/1996; Minsky, 1974; Simon, 1969). For example, a "danger" summary representation within the cognitive architecture of a squirrel would allow it, when traveling through three-dimensional space, to activate in real-time, episodes associated with the current location that have a "danger" tag, which would in turn inform decision-making about where to go and how careful to be (essentially, "Did I see a dog here before?"). Analogously, a "group" summary representation is what in principle allows an acute event (I punch Bob, who is a member of gang Z) to activate inferences aimed at a specific class of agents (I therefore now expect all members of gang Z to be angry at and retaliate against me; see also Pietraszewski & German, 2013).
6. Because the reader has a human brain, it will be exceedingly intuitive (almost redundant) to note that once we say that B is in a "group" with C but not A, such and such interaction type will occur. (That is, when told who is in a group with whom, the reader should feel the group-constitutive roles become obvious expectations of what will be more likely to happen.) This felt intuition is a prediction and occurs precisely because a group representation is, by hypothesis, these group-constitutive roles. So the fact that this is all very intuitive is good for the hypothesis. That said, for any readers who are bothered by the fact that spelling out the mental operational definition of a group necessarily violates the Gricean maxim of quantity, this section can be read as what one would need to program into a robot so that it, too, would have a concept of a social group in the context of conflict. [return to main text]
7. Ancillary attributes may also include mutually represented opportunities for establishing coordination and cooperation relationships (as in the "minimal" group paradigm; see Balliet, Wu, & De Dreu, 2014; Dunham, 2018; Pietraszewski, 2013, 2020). Describing these as *ancillary* is not to undermine the psychological importance of these attributes, but rather to articulate their precise role in the broader computational theory: that these activate group representations, but they are not the type itself.
8. The current proposal is also more explicit about what is done with these calculations: they determine which types of triadic interactions will be pulled up or activated as most likely – which is then consulted by other mechanisms for guiding behavior and generating predictions. That is, the current proposal is explicit about what functional outputs are being generated by the mechanism once "group" categories are represented – an issue upon which the containment metaphor is silent. Crucially, the categorization mechanism *doesn't do anything* after it categorizes, because there is nothing in the process of categorization itself that tells the mechanism what to do with the content that it has just categorized. So while categorization may be an important part of the information-processing story, it is far from adequate on its own. Therefore, even if it is a helpful way to conceptualize certain phenomena such as the accentuation of between-category differences and within-category similarities, the containment metaphor should not be confused for an adequate computational theory or explanation of those phenomena.
9. To be clear, these metaphors have led to the collection of important data and the creation of elegant and empirically robust dependent measures (e.g., the identity-fusion approach of Whitehouse, 2018). The point is simply that these metaphors (and dependent measures) should not be confused for a literal description of what is going on – either on the ground, or in the mind.
10. At first blush, the current account may seem to suggest that intentions and motivations don't matter. For example, that if somehow someone slipped and fell and blindly executed the behaviors corresponding to group-constitutive roles, then they would somehow be on equal footing with someone who executed the same behaviors because they care about their fellow group members. In fact, the current account does *not* treat these two scenarios as equivalent *because* of the current account, not in spite of it. Recall that the goal of the cognitive system for representing groups is to attend to the class of relationships that predict the occupation of group-constitutive roles. What motivations and intentions are within this framework are information-processing representations that allow the cognitive system to represent whether or not a one-time event is diagnostic of future events (Sell, 2011; Sell et al., 2017; Weiner, 1995). That is, why intentions matter in the first place – the *aboutness* of intentions – is made meaningful by the class of events to which they refer and render more probable. In this case, what makes a particular intention or motivation "genuinely" about groups is that it will lead agents to occupy the group-constitutive roles across all four interaction types, both now and in the future. For example, if an agent is simply aggressive or opportunistic, and thus occupies role C in *Defense*, this will not be seen as "genuine" group-based motivation precisely because it will not cause that agent to occupy group-constitutive roles with

respect to the same agent in the future. The current account thereby provides a causal account of where (conflict-based) group-based intentions and motivations come from in the first place, and an objective basis for determining (and predicting) whether a particular scenario will be considered by human intuition to be an example of a genuine group intention or motivation, or not.

11. Balliet et al.'s (2014) framework is probably best conceptualized as a set of modifiers to the polyadic event types, rather than being the event types themselves.

12. Direct reciprocity, at best, can only give us *Alliance*, but only if we stipulate that it is "positive" (i.e., an exchange of benefits) direct reciprocity between A and C. Worse still, reciprocity still does not give us all of the cases that we would want (i.e., it would leave us blind in many "group" situations) because in *Alliance*, C's behavior toward B is "positive direct reciprocity" toward A only if A's attacking B is a benefit to C, but this will not always be the case (see Pietraszewski, 2016; SOM). Next, indirect reciprocity can only give us two of the four interaction types: *Displacement* and *Defense*, both of which are "negative" (i.e., an exchange of costs) "indirect reciprocity" (and this is being very generous in that we are stipulating for free that an attack on B is somehow a cost on C, which then gives this account *Defense*, which technically we shouldn't grant). No account of reciprocity, direct or indirect, can account for *Generalization*. Finally, even for the subset of dynamics that the reciprocity account can redescribe, the account is far less powerful (at best) than the present framework (and at worst, doesn't work at all), because nowhere in the reciprocity account is the identity of the other third party relevant, unlike the present account. That is, if you and Bob are in a negative indirect reciprocity relationship, you are in that kind of relationship no matter who Bob is imposing a cost on. This means that the reciprocity account cannot handle the simple fact that who is on who's side depends on who else is involved. Therefore, if you built a robot using reciprocity, it would not switch sides as function of who is involved (i.e., it could not unite with a political outgroup member to fight against an invading nation, unlike a real human "robot").

References

Abrams, D., & Rutland, A. (2008). The development of subjective group dynamics. In S. R Levy, & M. Killen (Eds.*), Intergroup attitudes and relations in childhood through adulthood* (pp. 47–65). Oxford University Press.

Alexander, R. A. (1987). *The biology of moral systems*. Routledge.

Allport, G. W. (1924). *Social psychology*. Houghton Mifflin.

Allport, G. W. (1954/1958). *The nature of prejudice*. Doubleday Anchor.

Archer, J. (1988). *The behavioural biology of aggression*. Cambridge University Press.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, 140, 1556–1581.

Balliet, D., Tybur, J. M., & Van Lange, P. A. M. (2017). Functional interdependence theory: An evolutionary account of social situations. *Personality and Social Psychology Review, 21*(4):361–388, https://doi.org/10.1177/1088868316657965.

Barrett, H. C. (2015). *The shape of thought: How mental adaptations evolve*. Oxford University Press.

Bar-Tal, D. (Ed.). (2001). *Intergroup conflicts and their resolution: A social psychological perspective*. Taylor & Francis.

Bissonnette, A., Perry, S., Barrett, L., Mitani, J. C., Flinn, M., Gavrilets, S., & de Wall, F. B. M. (2015). Coalitions in theory and reality: A review of pertinent variables and processes. *Behaviour*, *152*, 1–56.

Block, N. (1997). Semantics, conceptual role. Routledge Encyclopedia of Philosophy. Retrieved from http://cogprints.org/232/1/199712005.html.

Böhm, R., Rusch, H., & Baron, J. (2018). The psychology of intergroup conflict: A review of theories and measures. *Journal of Economic Behavior & Organization*. https://doi.org/10.1016/j.jebo.2018.01.020.

Böhm, R., Rusch, H., & Gürerk, Ö. (2016). What makes people go to war? Defensive intentions motivate retaliatory and preemptive intergroup aggression. *Evolution and Human Behavior, 37*, 29–34.

Böhm, R., Rusch, H., & Baron, J. (2020). The psychology of intergroup conflict: A review of theories and measures. *Journal of Economic Behavior & Organization, 178,* 947–962.

Bornstein, G. (2003). Intergroup conflict: Individual, group, and collective interests. *Personality and Social Psychology Review, 7*, 129–145.

Brewer, M. B. (2001). Identity and conflict. In D. Bar-Tal (Ed.), *Intergroup conflicts and their resolution: A social psychological perspective* (pp. 125–143). Taylor & Francis.

Brown, D. E. (1991). *Human universals*. McGraw-Hill.

Brown, R. (2000). *Group processes*. Blackwell.

Burani, N., & Zwicker, W. S. (2003). Coalition formation games with separable preferences. *Mathematical Social Sciences, 45*, 27–52.

Byrne, R. W., & Whiten, A. (Eds.). (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Clarendon.

Campbell, A. (2015). Women's competition and aggression. In D. M. Buss (Ed.), *The handbook of evolutionary psychology, volume 2: Integrations* (pp. 684–703). Wiley.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioural Science, 3*, 14–25.

Caplow, T. (1959). Further development of a theory of coalitions in the triad. *American Journal of Sociology, 64*, 488–493.

Chase, I. D. (1985). The sequential analysis of aggressive acts during hierarchy formation: An application of the "jigsaw puzzle" approach. *Animal Behavior, 1985,* 86–100.

Chomsky, N. (1980). Rules and representations. *The Behavioral and Brain Sciences, 3*, 1–61.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology, 55,* 110–125.

Cikara, M., Van Bavel, J. J., Ingbretsen, Z. A., & Lau, T. (2017). Decoding "us" and "them": Neural representations of generalized group concepts. *Journal of Experimental Psychology: General, 5,* 621–631.

Daly, M. (2015). Interpersonal conflict and violence. In D. M. Buss (Ed.), *The handbook of evolutionary psychology, volume 2: Integrations* (pp. 669–683). Wiley.

Dawes, R. M., Van De Kragt, A. J. C., & Orbell, J. M. (1988). Not me or thee but we: The importance of group identity in eliciting cooperation in dilemma situations: Experimental manipulations. *Acta Psychologica, 68,* 83–97.

De Deo, S., Krakauer, D. C., & Flack, J. C. (2010). Inductive game theory and the dynamics of animal conflict. *PLoS Computational Biology, 6*(5), e1000782, https://doi.org/10.1371/journal.pcbi.1000782.

De Dreu, C. K. W., & Gross, J. (2019). Revisiting the form and function of conflict: Neurobiological, psychological, and cultural mechanisms for attack and defense within and between groups. *Behavioral and Brain Sciences, 42,* 1–66.

De Dreu, C. K. W., Gross, J., Méder, Z., Giffin, M., Prochazkova, E., Krikeb, J., & Columbus, S. (2016). In-group defense, out-group aggression, and coordination failures in intergroup conflict. *Proceedings of the National Academy of Sciences, 133,* 10524–10529.

De Dreu, C. K. W., & Van Vianen, A. E. M. (2001). Managing relationship conflict and the effectiveness of organizational teams. *Journal of Organizational Behavior, 22,* 309–328.

Dennett, D. (1969/2002). *Content and consciousness.* Routledge.

Dennett, D. (1988). Cognitive wheels: The frame problem in artificial intelligence. In Z. W. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial intelligence* (pp. 41–65). Ablex.

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112,* 281–299.

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin, 139*, 477–496.

Deutsch, M. (1962). Cooperation and trust: Some theoretical notes. In Manasa R. Jones(Ed.), *Nebraska symposium on Motivation* (pp. 275–320). University of Nebraska Press.

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Sciences, 22*, 780–793.

Etzioni, O. (2018). *Learning common sense: A grand challenge for academic AI research.* Talk at the Office of Naval Research.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23,* 1–74.

Forsyth, D. R. (2014/2019). *Group dynamics* (7th ed.). Cengage.

Gamson, W. A. (1961). A theory of coalition formation. *American Sociological Review, 26*, 373–382.

Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science, 41,* 545–575.

Gershman, S. J., & Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science, 29*(5), 460–466.

Glowacki, L., Wilson, M. L., & Wrangham, R. W. (2020). The evolutionary anthropology of war. *Journal of Economic Behavior & Organization, 178,* 963–982.

Goldstone, R. L., Roberts, M. E., & Gureckis, T. M. (2008). Emergent processes in group behavior. *Current Directions in Psychological Science, 17*, 10–15.

Gonzalez, C., Ben-Asher, N., Martin, J. M., & Dutt, V. (2015). A cognitive model of dynamic cooperation with varied interdependency information. *Cognitive Science, 39*, 457–495.

Grammer, K. (1992). Intervention in conflicts among children: Contexts and consequences. In A. H. Harcourt, & F. B. M. de Waal (Eds.), *Coalitions and alliances in humans and other animals* (pp. 259–283). Oxford University Press.

Gross, J., & De Dreu, C. K. W. (2019). The rise and fall of cooperation through reputation and group polarization. *Nature Communications, 10,* 776 https://doi.org/10.1038/s41467-019-08727.

Hamilton, D. L., Sherman, S. J., & Castelli, L. (2002). A group by any other name – The role of entitativity in group perception. *European Review of Social Psychology, 12,* 139–166.

Harcourt, A. H. (1988). Alliances in contests and social intelligence. In R.W. Byrne, & A. Whiten (Eds.), *Machiavellian intelligence* (pp. 132–152). Clarendon Press.

Harcourt, A. H., & deWall, F. B. M. (Eds.). (1992). *Coalitions and alliances in humans and other animals.* Oxford University Press.
Hardy, I. C. W., & Briffa, M. (Eds.). (2013). *Animal contests.* Cambridge University Press.
Haslam, S. A., Reicher, S. D., & Platow, M. J. (2011). *The new psychology of leadership: Identity, influence and power.* Psychology Press.
Heider, F. (1958). *The psychology of interpersonal relations.* Erlbaum.
Insko, C. A., & Schopler, J. (1987). Categorization, competition and collectivity. In C. Hendrick (Ed.), *Group processes* (Vol. 8, pp. 213–251). Sage.
Ip, G. W., Chiu, C., & Wan, C. (2006). Birds of a feather flock together: Physical versus behavioral cues may lead to trait- versus goal-based group perception. *Journal of Personality and Social Psychology, 90*, 368–381.
Jackendoff, R. S. (1992). *Languages of the mind: Essays on mental representation.* MITPress.
Kamermans, M., & Schmits, T. (2004). The history of the frame problem. *Artificial Intelligence, 86,* 116.
Koykka, C., & Wild, G. (2017). Concessions, lifetime fitness consequences, and the evolution of coalitionary behavior. *Behavioral Ecology, 28,* 20–30.
Lau, T., Gershman, S. J., & Cikara, M. (2020). Social structure learning in human anterior insula. *eLife, 9,* e53162.
Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General, 147,* 1881–1891.
Levine, R., & Campbell, D. (1972). *Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior.* Wiley.
Lewin, K. (1948). *Resolving social conflict.* Harper.
Lorenz, K. (1948/1996). *The natural science of the human species: An introduction to comparative behavioral research, the "Russian Manuscript"* (1944–1948). MIT Press.
Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology, 78,* 223–246.
Liska, G. (1962). *Nations in alliance: The limits of interdependence.* Johns Hopkins University Press.
Lopez, A. C. (2017). The evolutionary psychology of war: Offense and defense in the adapted mind. *Evolutionary Psychology, 15,* 1–23. https://doi.org/10.1177/1474704917742720.
Lopez, A. C. (2020). Making 'my' problem 'our' problem: Warfare as collective action, and the role of leader manipulation. *The Leadership Quarterly, 31*(2), 101294.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* W H Freeman.
Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B, 200,* 269–294.
McGrath, J. E. (1984). *Groups: Interaction and performance.* Prentice-Hall.
Mesterton-Gibbons, M., Gavrilets, S., Gravner, J., & Akcay, E. (2011). Models of coalition or alliance formation. *Journal of Theoretical Biology, 274*(1), 187–204.
Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE,* 49, 8–30.
Minsky, M. (1974). A framework for representing knowledge. Artificial Intelligence Memo No. 306.
Moncrieff, M., & Lienard, P. (2019). What war narratives tell about the psychology and coalitional dynamics of ethnic violence. *Journal of cognition and culture, 19*(1–2), 1–38.
Mueller, E. T. (2015). *Commonsense reasoning: An event calculus based approach.* Morgan Kaufmann.
Oliver, P. E. (1993). Formal models of collective action. *Annual Review of Sociology, 19,* 271–300.
Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences, 21*(6), 425–433.
Park, J. H., & Van Leeuwen, F. (2015). Evolutionary perspectives on social identity. In V. Zeigler-Hill, L. Welling, & T. Shackelford (Eds.), *Evolutionary perspectives on social psychology* (pp. 115–125). Springer.
Patton, J. Q. (1996). Thoughtful warriors: Status, warriorship, and alliance in the Ecuadorian Amazon. PhD dissertation, University of California Santa Barbara.
Patton, J. Q. (2000). Reciprocal altruism and warfare: A case from the Ecuadorian Amazon. In L. Cronk, N. Chagnon, & W. Irons (Eds.), *Adaptation and human behavior: An anthropological perspective* (pp. 417–436). Aldine de Gruyter.
Peirce, C. S. (1931–1935). Collected papers of Charles Sander Peirce. In P. Weiss, C. Hartshorne, & A. W. Burks (Eds.). Harvard University Press.
Perc, M., & Szolnoki, A. (2010). Coevolutionary games – A mini review. *BioSystems, 99,* 109–125.
Perry, S., Barrett, H. C., & Manson, J. H. (2004). White-faced capuchin monkeys show triadic awareness in their choice of allies. *Animal Behaviour, 67*, 165–170.
Pietraszewski, D. (2012). The elementary dynamics of intergroup conflict and revenge. *Behavioral and Brain Sciences, 36,* 32–33.
Pietraszewski, D. (2013). What is group psychology? Adaptations for mapping shared intentional stances. In M. Banaji, & S. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 253–257). Oxford University Press.
Pietraszewski, D. (2016). How the mind sees coalitional and group conflict: The evolutionary invariances of n-person conflict dynamics. *Evolution and Human Behavior, 37,* 470–480.
Pietraszewski, D. (2020). Intergroup processes: Principles from an evolutionary perspective. In P. Van Lange, E. T. Higgins, & A. W. Kruglanski (Eds.), *In social psychology: Handbook of basic principles* (pp. 373–391). Guilford.
Pietraszewski, D., Curry, O., Peterson, M. B., Cosmides, L., & Tooby, J. (2015). Constituents of political cognition: Race, party politics, and the alliance detection system. *Cognition*, *140,* 24–39.
Pietraszewski, D., & German, T. C. (2013). Coalitional psychology on the playground: Reasoning about indirect social consequences in preschoolers and adults. *Cognition, 126,* 352–363.
Pietraszewski, D., & Wertz, A. E. (2021). Why evolutionary psychology should abandon modularity. *Perspectives on Psychological Science,* 1–26. doi: 10.1177/1745691621997113.
Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Science.* doi: 201304326.
Puurtinen, M., Heap, S., & Mappes, T. (2015). The joint emergence of group competition and within-group cooperation. *Evolution and Human Behavior, 36*, 211–217.
Rhodes, M., & Chalik, L. (2013). Social categories as markers of intrinsic interpersonal obligations. *Psychological Science, 24,* 999–1006.
Rusch, H. (2013). Asymmetries in altruistic behavior during violent intergroup conflict. *Evolutionary Psychology, 11,* 973–993.
Rusch, H. (2014a). The evolutionary interplay of intergroup conflict and altruism in humans: A review of parochial altruism theory and prospects for its extension. *Proceedings of the Royal Society B, 281,* 20141539.
Rusch, H. (2014b). The two sides of warfare. *Human Nature, 25,* 359–377.
Rusch, H., & Gavrilets, S. (2020). The logic of animal intergroup conflict: A review. *Journal of Economic Behavior & Organization, 178,* 1014–1030.
Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach.* Prentice Hall.
Sandholm, T., Larson, K., Andersson, M., Sherry, O., & Tohmé, F. (1999). Coalition structure generation with worst case guarantees. *Artificial Intelligence, 111,* 209–238.
Schelling, T. C. (1956). An essay on bargaining. *The American Economic Review, 46,* 281–306.
Schelling, T. C. (1966). *The strategy of conflict.* Harvard University Press.
Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate-proximate distinction in the human behavioral sciences. *Perspectives in Psychological Science, 6,* 38–47.
Sedikides, C., Schopler, J., & Insko, C. A. (Eds.). (1998). *Intergroup cognition and intergroup behavior.* Erlbaum.
Sell, A. N. (2011). The recalibrational theory and violent anger. *Aggression and Violent Behavior, 16,* 381–389.
Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., … Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition, 168,* 110–128.
Shaw, A., DeScioli, P., Barakzai, A., & Kurzban, R. (2017). Whoever is not with me is against me: The costs of neutrality among friends. *Journal of Experimental Social Psychology, 71,* 96–104.
Shaw, M. E. (1981). *Group dynamics: The psychology of small group behavior* (3rd ed.). McGraw-Hill.
Shehory, O., & Kraus, S. (1998). Methods for task allocation via agent coalition formation. *Artificial Intelligence, 101,* 165–200.
Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review, 1,* 2–28.
Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. (1961). *Intergroup conflict and cooperation: The robbers' cave experiment.* University of Oklahoma Press.
Sherratt, T. M., & Mesterton-Gibbons, M. (2013). Models of group or multi-party contests. In I. C. W. Hardy, & M. Briffa (Eds.), *Animal contests* (pp. 33–46). Cambridge University Press.
Simon, H. A. (1969/1996). *The sciences of the artificial* (3rd ed.). MIT Press.
Smaldino, P., Pickett, C., Sherman, J., & Schank, J. (2012). An agent-based model of social identity dynamics. *Journal of Artificial Societies and Social Simulation, 15,* 7.
Strayer, F. F., & Noel, J. M. (1986). The prosocial and antisocial functions of preschool aggression. In C. Zahn-Waxler, E. M. Cummings, & R. Iannotti (Eds.), *Altruism and aggression: Biological and social origins* (pp. 107–131). Cambridge University Press.
Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American, 223,* 96–102.
Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology, 33,* 1–39.
Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin, & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.
Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel, & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Nelson-Hall.

Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition, 137,* 47–62.
Tooby, J., & Cosmides, L. (1992). The cognitive foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). Oxford Press.
Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory.* Blackwell.
Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences, 21*(9), 649–665.
Vainstein, M. H., Silva, A. T., & Arenzon, J. J. (2007). Does mobility decrease cooperation? *Journal of Theoretical Biology, 244,* 722–728.
Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton University Press.
Walter, B. F. (2004). Does conflict beget conflict? Explaining recurring civil war. *Journal of Peace Research, 41,* 371–388.
Walter, B. F. (2009). *Reputation and civil war.* Cambridge University Press.
Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* Guilford Press.
Whitehouse, H. (2018). Dying for the group: Towards a general theory of extreme self-sacrifice. *Behavioral and Brain Sciences, 41,* 1–64.
Wilder, D., & Simon, A. F. (1998). Categorical and dynamic groups: Implications for social perception and intergroup behavior. In C. Sedikides, J. Schopler, & C. A. Insko (Eds.), *Intergroup cognition and intergroup behavior* (pp. 27–44). Erlbaum.
Worden, L., & Levin, S. A. (2007). Evolutionary escape from the prisoner's dilemma. *Journal of Theoretical Biology, 245,* 411–422.
Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity: Ingroup boasting and ingroup favoritism. *Advances in Group Processes, 16,* 161–197.

Open Peer Commentary

Using laboratory intergroup conflict and riots as a "stress test"

James M. Allen and Daniel C. Richardson
Department of Psychology, University College London (UCL), London WC1H 0AP, UK
jmallen@cantab.net | daniel.richardson@ucl.ac.uk | https://www.ucl.ac.uk/pals/research/experimental-psychology/person/daniel-richardson/
| doi:10.1017/S0140525X21001333, e98

Abstract

We apply the author's computational approach to groups to our empirical work studying and modelling riots. We suggest that assigning roles in particular gives insight, and measuring the frequency of bystander behaviour provides a method to understand the dynamic nature of intergroup conflict, allowing social identity to be incorporated into models of riots.

1. Introduction

We intend to "stress-test" the author's computational approach to groups using concrete examples of intergroup conflict – specifically, riots. Drawing on psychological and sociological accounts of the London riots of 2011, we previously developed a series of lab-based games of collective action where we experimentally manipulate psychology factors and use agent-based modelling to understand our participants' behaviour. We consider the idea proposed in the article that individuals perceive a group through roles and behaviour, and examine whether this gives insight into our findings.

2. Application of the framework to our empirical investigations

Our experimental paradigm (Dezecache, Allen, von Zimmerman, & Richardson, 2021) tests conflict between groups with unequal resources, and asks how relative deprivation and social identity interact to generate conflict in the laboratory. Players are assigned to two teams in a computer game played in person or online. They tap on their devices to build features in a park, such as a flower bed. However, players also have the option to vandalise the other team's park, trampling the flower beds or damaging the playground, and assign a disadvantage to one of the teams – they have to work harder to build their park features. We find that this structural inequality in the game produces a higher rate of aggressive behaviour towards the advantaged team, a conclusion supported through our agent-based models.

Within our paradigm, groups can also be a mix of preexisting identities and minimal group assignments, allocated in different proportions across conditions. Therefore, although players are assigned at random to the advantaged or disadvantaged park, they are also informed, for example, that their team mates share their political beliefs, or have varying political opinions. An open question is then whether individuals consider others within their empirically assigned team to be part of their group, or if it is preexisting identities driving any observed intergroup conflict.

Considering the target article's computational approach, we find that insight is gained by assuming that participants compute groups through assigning roles in triads, and particularly through considering the frequency of observed behaviours between different groups. By observing the behaviour of the bystander, and taking the assumption that participants compute group membership through the observation of roles and behaviour, we could infer which groups participant's consider themselves to be part of through observed triad frequencies. As we observe changing proportions of vandalism over time, the frequency of different triads may also inform which group (e.g., team or wider identity) is salient at any one time. By observing the behaviour of different bystanders towards the relative success of others, we may be able to better understand how individuals are making these comparisons, and whether groups or individual comparisons are key. Therefore, depending on the level of observed conflict between groups, we can infer if participants are considering individual or group relative deprivation.

However, the fuzzy nature of real group dynamics means there are a number of things we cannot infer by assuming that individuals use this computational method. Our own modelling points towards coordination and conflict coexisting within and between groups. By defining groups in a concrete manner, the opportunity for a wider number of roles is lost. If an individual performs a different task to the rest of the group, does that mean they are not a member of the group, or does it just mean that they are a part of the group, but coordinating behaviour?

Finally, to infer feelings of relative deprivation, and also any other drivers of intergroup conflict, we feel that more attention must be paid to an agent's internal state. Although assigning roles based on behaviour works in areas such as conflict, which is an overt behaviour, this is not true of affective states such as frustration, or even the act of making a comparison. Therefore, how is it that an observer could then classify others into groups?

We also consider how this computational approach contributes to our understanding of real-world riots. One explanation from social psychology is that riots emerge through aggressive action perceived through salient identities (e.g., Drury et al., 2020; Reicher, 1984; Stott & Drury, 2017; Stott et al., 2018). This explanation aligns with the target article's framework. Rather than defining groups through a number of context-dependent external markers, it is the role the individual takes within this scenario that is important. By understanding conflict through the allocation of roles, this framework also introduces a flexibility to help us understand this phenomenon across scales. What begins as a local phenomenon can soon spread, as an individual bystander observing aggression between two individuals or two groups is spurred into action in a similar way.

3. Conclusion

We find the idea of asking what is the minimum information required to understand a "group" intriguing. When applied to the real-world example of riots, this framework provides insight through its flexibility to describe conflict across scales, and its ability to marry contagion models with social psychology. Within our empirical paradigm, by assuming that participants do indeed compute groups in this way, we may gain information on exactly who they consider to be part of their group.

More usefully, this framework adds to the continuing discussion on how best to model a riot (e.g., Baudains, Braithwaite, & Johnson, 2013a; Baudains, Johnson, & Braithwaite, 2013b; Bonnasse-Gahot et al., 2018; Davies, Fry, Wilson, & Bishop, 2013). Currently, most models consider contagion to be the central mechanism of riot spreading. However, what is lacking from these models is the inclusion of the importance of the identity salience to who imitates who, how this changes over time, and how this structures their social interactions. However, by including bystanders, and considering how they are likely to act next, the computational framework enables us to understand why these contagion models are successful at matching the data, even before many of the factors discussed in social psychology are considered.

However, despite the simplicity and flexibility of this framework, it does have a number of issues. Specifically, we find that the myriad of possible behaviours required to explain more complicated real-world scenarios, such as we find in the laboratory, cannot be understood solely through this computational model.

Financial support. JMA and DCR are funded by a grant from the Nuffield Foundation (The Psychological Roots of Societal Self Harm, 42868) awarded to DCR.

Conflict of interest. None.

References

Baudains, P., Braithwaite, A., & Johnson, S. D. (2013a). Target choice during extreme events: A discrete spatial choice model of the 2011 London riots. *Criminology; An interdisciplinary Journal, 51,* 251–285.

Baudains, P., Johnson, S. D., & Braithwaite, A. M. (2013b). Geographic patterns of diffusion in the 2011 London riots. *Applied Geography, 45,* 211–219.

Bonnasse-Gahot, L., Berestycki, H., Depuiset, M.-A., Gordon, M. B., Roché, S., Rodriguez, N., & Nadal, J.-P. (2018). Epidemiological modelling of the 2005 French riots: A spreading wave and the role of contagion. *Scientific Reports, 8,* 107. doi: 10.1038/s41598-017-18093-4

Davies, T. P., Fry, H. M., Wilson, A. G., & Bishop, S. R. (2013). A mathematical model of the London riots and their policing. *Scientific Reports, 3,* 1303.

Dezecache, G., Allen, J. M., von Zimmerman, J., & Richardson, D. C. (2021). We predict a riot: Inequity, relative deprivation and collective destruction in the laboratory. *Proceedings of the Royal Society B, 288,* 20203091. doi: 10.1098/rspb.2020.3091

Drury, J., Stott, C., Ball, R., Reicher, S., Neville, F., Bell, L., & Ryan, C. (2020). A social identity model of riot diffusion: From injustice to empowerment in the 2011 London riots. *European Journal of Social Psychology.* n/a. doi: doi.org/10.1002/ejsp.2650

Reicher, S. D. (1984). The St. Pauls' riot: An explanation of the limits of crowd action in terms of a social identity model. *European Journal of Social Psychology, 14*, 1–21. doi: https://doi.org/10.1002/ejsp.2420140102

Stott, C., Ball, R., Drury, J., Neville, F., Reicher, S., Boardman, A., & Choudhury, S. (2018). The evolving normative dimensions of "riot": Towards an elaborated social identity explanation. *European Journal of Social Psychology, 48,* 834–849. doi: 10.1002/ejsp.2376

Stott, C., & Drury, J. (2017). Contemporary understanding of riots: Classical crowd psychology, ideology and the social identity approach. *Public Understanding of Science, 26,* 2–14. doi: 10.1177/0963662516639872

Beyond folk-sociology: Extending Pietraszewski's model to large-group dynamics

Pascal Boyer

Department of Psychology and Brain Sciences, Washington University in St. Louis, St. Louis, MO 63130, USA
pboyer@wustl.edu | http://www.pascalboyer.net | doi:10.1017/S0140525X21001278, e99

Abstract

Folk-sociology is a set of intuitive assumptions that organize our spontaneous theories about society, including the notion that social groups are agent-like. Pietraszewski's model may explain this folk-sociological assumption in an elegant way. However, large-scale group dynamics include features that seem to escape agent-like descriptions. Therefore, one may want to find out whether the "event-grammar" proposed here can account for these features.

Pietraszewski's proposal for the cognitive underpinnings of the "social group" concept is a remarkable attempt to place social science concepts on a lucid foundation, in this case to account for people's own concepts of their social environment in a computationally tractable manner. This good deed should not go unpunished, so I will argue for a major extension of the program.

Pietraszewski's model requires that we abandon common, entrenched intuitions about collections of agents and their interaction. These intuitions are not just the product of past social theory, they also constitute the way most lay people, in the most diverse cultures, construe the social world – what could be called a *folk-sociology* (Boyer, 2018, pp. 216–237; Hirschfeld, 2001). For instance, one common assumption is that social norms are somehow external to the individual minds that represent them, which, for example, makes it possible to think that, for example, "marriage *is*…" this or that, regardless of people's thoughts about it. Another important and culturally widespread assumption is that groups are agent-like entities, which is why we talk about villages or social classes or nations as entities that want this or remember that, make decisions, and so on. Finally, folk-sociology assumes that political power is a "force" and power relations are similar to force-dynamics. Powerful exert "pressure" on others, who may "resist" or be "pushed around," and so on (Boyer, 2018).

These assumptions are based on loose and misleading conventional metaphors (Lakoff & Johnson, 1980). Therefore, why are they so widespread and entrenched? A first reason is that they describe complex processes of social interaction, with emergent properties, for which we do not have adequate cognitive resources. Even in the small-scale groups typical of most human evolution, describing interaction between more than three individuals would require not just representing other agents' intentions, but their representations of each others' representations, and so forth. Therefore, folk-sociological assumptions produce a rough and ready understanding of social processes that may be sufficient in many situations. A second reason that makes folk-sociological assumptions quite natural is that they often "hijack" the computational machinery of evolved domain-specific cognitive systems. Seeing groups as agents makes it possible to activate our intuitive psychology (Leslie, Friedman, & German, 2004), which provides a rich descriptive and explanatory arsenal for describing group interactions. As regards power, intuitive understandings of force dynamics (Baillargeon, Kotovsky, & Needham, 1995; Talmy, 1988) supply templates for describing the dynamics of influence between leaders and followers. A third factor is that other agents too share our folk-sociological assumptions, so that these partly metaphorical understandings provide coordination points, a form of mutual knowledge about our social environment.

In a very unfortunate turn of events, a great deal of social science theorizing, instead of explaining folk-sociology, endorsed it! Therefore, that social scientists tried to argue that norms really were external to people's minds (Gilbert, 1989; Searle, 1995), that power really worked like a force – see discussion in Lukes (1984), and that social groups really had intentional states (Sheehy, 2012; Tollefsen, 2015).

Dave Pietraszewski's model may help explain the groups-as-agents assumption of folk-sociology. Groups are seen as intentional, having both goals and memories, not just (as suggested above) because our intuitive psychology is a salient mental system with a rich inferential potential, but also because the very notion of social group requires an intentional description of the "primitive roles" that constitute non-cooperative triadic interactions. In other words, the top-down salient metaphor of groups as agents may be salient because of the actual bottom-up process of considering (within the triadic interactions) agents as constitutive of groups. It is agency all the way down, so to speak.

But one may want to know how this parsimonious model accommodates some aspects of large-scale group dynamics. Consider this. At the level of actual triadic interactions, or perhaps an extension to a few individuals in each "group-defining role," the participants' representations of the different roles and their consequences are mostly veridical. That is, people do represent the imposition of cost as such, and the reactions of the other partners as one of the roles defined in the event grammar. By contrast, the dynamics of interaction at a larger scale may be quite different from these roles and depend on factors not described by the event-grammar. Consider for instance the dynamics of ethnic signaling, as modeled by Kuran (1998). People's decision to wear ethnic garb, for instance, depends on the perceived cost of doing so, which, in turn, depends on an agent's perception of the relative size and cohesion of one's and the other groups. The decision, in turn, influences other agents, as it modifies their estimates of the relative costs of signaling or not signaling. Naturally, this cascade model is not at all the way people represent the macro-features of the situation – which they see as people beginning to signal their ethnicity because they are more convinced than before of their group's value and of its goals.

There are two possible paths to account for such situations. Pietraszewski suggests that the event-grammar scales up, preserving its constitutive roles. In this case, people's own representations of large-scale group dynamics do capture people's actual behavior. The macro-effects outlined here are adequately described as the emergent effect of numerous agents entertaining the representations described by the event-grammar.

Alternatively, we may consider that large-scale group dynamics cannot be captured by the event-grammar. In that case, the macro-dynamics are really different from what people represent. But the fact that people all see those dynamics in terms of agent-like groups provides enough coordination that the intentional description often appears to be roughly correct so that most agents (and even social scientists!) see it as adequate.

It speaks to the great value and conceptual precision of Pietraszewski's proposal, that we can formulate these questions with a level of precision that is not common in the social sciences.

Conflict of interest. None.

References

Baillargeon, R., Kotovsky, L., & Needham, A. (1995). The acquisition of physical knowledge in infancy. In D. Sperber, D. Premack, & A. James-Premack (Eds.), *Causal cognition. A multidisciplinary debate* (pp. 79–115). Clarendon Press.
Boyer, P. (2018). *Minds make societies. How cognition explains the world humans create.* Yale University Press.
Gilbert, M. (1989). *On social facts.* Routledge.
Hirschfeld, L. A. (2001). On a folk theory of society: Children, evolution, and mental representations of social groups. *Personality and Social Psychology Review, 5,* 107–117. doi: 10.1207/S15327957PSPR0502_2.
Kuran, T. (1998). Ethnic norms and their transformation through reputational cascades. *Journal of Legal Studies, 27,* 623–659.
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* The University of Chicago Press.
Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind." *Trends in Cognitive Sciences, 8,* 529–533.
Lukes, S. (1984). *Power: A radical view.* Macmillan Press.
Searle, J. R. (1995). *The construction of social reality.* Free Press.
Sheehy, P. (2012). *The reality of social groups:* Ashgate.
Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12,* 49–100. 10.1016/0364-0213(88)90008-0.
Tollefsen, D. P. (2015). *Groups as agents.* Wiley.

Signals and cues of social groups

Gregory A. Bryant and Constance M. Bainbridge

Department of Communication, Center for Behavior, Evolution, and Culture, University of California, Los Angeles, Los Angeles, CA 90095-1563, USA gabryant@ucla.edu; http://gabryant.bol.ucla.edu/ | cbainbridge@ucla.edu; http://constancebainbridge.com/ |

doi:10.1017/S0140525X21001461, e100

Abstract

A crucial factor in how we perceive social groups involves the signals and cues emitted by them. Groups signal various properties of their constitution through coordinated behaviors across sensory modalities, influencing receivers' judgments of the group and subsequent interactions. We argue that group communication is a necessary component of a comprehensive computational theory of social groups.

Pietraszewski provides a compelling computational framework for understanding how people represent and reason about social groups, but the approach requires logical extensions. For example, the model does not factor in the nature of bonds between ingroup members, and how those associations might be communicated to others. We should expect machinery for processing signals of groups that could create or update existing representations. Social connections can influence a group's perceived entitativity, and consequently how others will potentially interact with it (e.g., whether to retaliate on another's behalf or not). Research has revealed possible strategies of how groups can either collectively signal their existence, or reveal themselves through by-product cues arising in different modalities. Here, we will describe some recent research exploring acoustic signals and cues of affiliation and group membership.

The question of how people might detect groups based on observable behavior requires making the distinction between adaptive signals and by-product cues (Maynard Smith & Harper, 2003). Signals are communicative adaptations designed to affect the behavior of other organisms, and evolve in tandem with receiver adaptations. Signaling systems generally evolve by conferring mutual benefits to senders and receivers on average. Conversely, cues are detectable by-products of a phenotype that are not designed to convey information, but they can nevertheless shape receiver perceptual systems. Cost–benefit trade-offs underlie the dynamics of whether cues are maintained by selection; that is, they are typically associated with necessary components of physiological, cognitive, or behavioral systems that serve other purposes. For example, deceptive signaling is often associated with a variety of effects that are difficult for senders to conceal because of cognitive load and/or emotional effects (DePaulo et al., 2003).

Most generally, observable social interactions provide many cues of group structure. Group detection is a flourishing area of research in computer vision, revealing detectable structure in small groups in crowds (Wang, Chen, Nie, & Li, 2018), conversationalists (Vascon et al., 2014), and even rapport between interlocutors (Müller, Huang, & Bulling, 2018), among others. Definitions of groups by researchers in this domain typically suffer from the vagueness described by Pietraszewski. Nevertheless, correlates of social group behavior are often not present by design but as by-products of social interaction patterns. Other markers are by design however, and will culturally evolve if certain social ecological conditions are met related to the evolutionary stability of cooperative interaction between the relevant individuals (McElreath, Boyd, & Richerson, 2003).

In the domain of auditory communication, a variety of possible signals of social groups have been examined, including nonlinguistic vocalizations such as colaughter, linguistic phenomena such as speech accents, and group musical behavior. For instance, listeners across two dozen disparate societies were able to reliably detect friends and strangers from very brief (<2 s) recordings of dyadic colaughter (Bryant et al., 2016). Colaughter is better than cospeech at revealing affiliation (Bryant, Wang, & Fusaroli, 2020), and infants as young as 5-months are sensitive to this information (Vouloumanos & Bryant, 2019). Moreover, the acoustic features of human colaughter implicate a broadcast function as it is often loud, abrupt, conspicuous, and repetitive (Bryant et al., 2020).

Production and perceptual data suggest that colaughter could constitute a signal of affiliation – one that often operates beyond the awareness of the signalers. Group interactions can not only result in behaviors that function to communicate group structure without the knowledge of the signalers, but also evolutionary processes can shape communication systems to have features that ultimately serve to cue social categories. For example, speech accents arise when adult second-language learners fail to acquire phonological and morphosyntactic idiosyncrasies of a given language. Accents are highly detectable, and are a fixed aspect of social categorization, not a by-product of coalitionary reasoning (Pietraszewski & Schwartz, 2014a, 2014b).

Another domain where group signaling occurs is that of music. Human music is plausibly linked evolutionarily to the coordinated vocalizations of many species signaling territory boundaries (Hagen & Bryant, 2003; Hagen & Hammerstein, 2009), and rooted in emotional signaling and the coordination of group affect (Bryant, 2013, 2014). One important mechanism for coordinating signals effectively is physical entrainment, allowing for social groups to collectively produce rhythmic displays (Phillips-Silver, Aktipis, & Bryant, 2010; Ravignani, Bowling, & Fitch, 2014). Rhythmic production and linked perception-action mechanisms potentially constitute core musical adaptations (among possible others) that underlie the cultural evolution of music, affording complex social group signaling at multiple levels (Mehr, Krasnow, Bryant, & Hagen, 2021).

Auditory communication can be an especially effective medium for groups to advertise. Many nonhuman animals exploit the sonic environment for this purpose, including social carnivores and primates (Hagen & Hammerstein, 2009). For example, a single covocalizing pair of wolves has been shown to be perceived as a larger group by listeners (Harrington, 1989), a phenomenon coined the *Beau Geste* effect (Krebs, 1977). Currently, in our lab we are exploring differences across human vocal modalities (e.g., laughter, yelling, and speaking) in how listeners judge group size, that could point to possible vocal adaptations in humans for communicating about social groups. Early results suggest, unsurprisingly, that yelling affords perceptions of larger group sizes, and that colaughter seems to reduce size estimates, making groups sound more integrated. These differences may reveal distinct social signaling functions of varying kinds of human covocalizations and could be understood in the context of the computational framework presented by Pietraszewski. For example, if collective yelling revealed strength in defending a resource, bystanders may engage such a group with a primed expectation of conflict, and assess risk appropriately. Alternatively, if affiliative laughter signals a different social ecology, it could encourage others to more readily join a group, given the relatively lowered threat. Particular patterns of signaling may often characterize specific role arrangements in group interactions, with the associations reliably learned in social agents.

We propose that a thorough treatment of social signaling across all communicative channels is required for a comprehensive computational theory of social groups, and that signaling adaptations must be conceptually separated from the reliable production and detection of by-product cues. We have provided some preliminary examples of recent research in vocal communication and the evolution of music that point in this direction.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. Neither author reports any conflicts of interest.

References

Bryant, G. A. (2013). Animal signals and emotion in music: Coordinating affect across groups. *Frontiers in Psychology, 4,* 990, 1–13.

Bryant, G. A. (2014). The evolution of coordinated vocalizations before language. *Behavioral and Brain Sciences, 37*(6), 549–550.
Bryant, G. A., Fessler, D. M. T., Fusaroli, R., Clint, E., Aarøe, L., Apicella, C. L., … Zhou, Y. (2016). Detecting affiliation in colaughter across 24 societies. *Proceedings of the National Academy of Sciences, 113*(17), 4682–4687.
Bryant, G. A., Wang, C. S., & Fusaroli, R. (2020). Recognizing affiliation in colaughter and cospeech. *Royal Society Open Science, 7*(10), 201092.
DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74–118.
Hagen, E. H., & Bryant, G. A. (2003). Music and dance as a coalition signaling system. *Human Nature, 14*(1), 21–51.
Hagen, E. H., & Hammerstein, P. (2009). Did Neanderthals and other early humans sing? Seeking the biological roots of music in the territorial advertisements of primates, lions, hyenas, and wolves. *Musicae Scientiae, 13*(2_suppl), 291–320.
Harrington, F. H. (1989). Chorus howling by wolves: Acoustic structure, pack size and the Beau Geste effect, *Bioacoustics, 2*(2), 117–136.
Krebs, J. R. (1977). The significance of song repertoires: The Beau Geste hypothesis. *Animal Behaviour, 25*(2), 475–478.
Maynard Smith, J., & Harper, D. (2003). *Animal signals.* Oxford University Press.
McElreath, R., Boyd, R., & Richerson, P. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology, 44*(1), 122–130.
Mehr, S., Krasnow, M., Bryant, G. A., & Hagen, E. H. (2021). Origins of music in credible signaling. *Behavioral and Brain Sciences, 44,* e60. doi:10.1017/S0140525X20000345
Müller, P., Huang, M. X., & Bulling, A. (2018). Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *The 23rd international conference on intelligent user interfaces* (pp. 153–164). Available at: https://doi.org/10.1145/3172944.3172969.
Phillips-Silver, J., Aktipis, A., & Bryant, G. A. (2010). The ecology of entrainment: Foundations of coordinated rhythmic movement. *Music Perception, 28*(1), 3–14.
Pietraszewski, D., & Schwartz, A. (2014a). Evidence that accent is a dimension of social categorization, not a byproduct of perceptual salience, familiarity, or ease-of-processing. *Evolution and Human Behavior, 35*(1), 43–50.
Pietraszewski, D., & Schwartz, A. (2014b). Evidence that accent is a dedicated dimension of social categorization, not a byproduct of coalitional categorization. *Evolution and Human Behavior, 35*(1), 51–57.
Ravignani, A., Bowling, D. L., & Fitch, W. (2014). Chorusing, synchrony, and the evolutionary functions of rhythm. *Frontiers in Psychology, 5,* 1118.
Vascon, S., Mequanint, E. Z., Cristani, M., Hung, H., Pelillo, M., & Murino, V. (2014). A game-theoretic probabilistic approach for detecting conversational groups. In *Asian conference on computer vision* (pp. 658–675). Springer.
Vouloumanos, A., & Bryant, G. A. (2019). Five-month-old infants detect affiliation in colaughter. *Scientific Reports, 9*(1), 1–8.
Wang, Q., Chen, M., Nie, F., & Li, X. (2018). Detecting coherent groups in crowd scenes by multiview clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(1), 46–58.

Latent structure learning as an alternative computation for group inference

Mina Cikara

Psychology Department, Harvard University, Cambridge, MA 02138, USA mcikara@fas.harvard.edu |
https://www.intergroupneurosciencelaboratory.com | doi:10.1017/S0140525X21001254, e101

Abstract

In contrast to Pietraszewski’s account, latent structure learning neither requires conflict nor relies on observation of explicit coalitional behavior to support group inference. This alternative addresses how even non-conflict-based groups may be defined and is supported by experimental evidence in human behavior.

Pietraszewski presents a computational theory of how people solve the important problem of figuring out what constitutes a social group in the context of conflict. I agree that past definitions of “groups” are in many cases circular and insufficient (Cikara & Van Bavel, 2014). We are long overdue for a high-level formalization of how people solve the “group” problem and I am grateful to Pietraszewski for initiating this conversation. In my view, however, the target article’s theory faces two major challenges. First, though Pietraszewski speculates about non-conflict-based group representations, the theory still falls short of offering a unified account of group representations across different coalitional contexts (e.g., when agents are *not* in conflict). Second, though the stated goal of this paper is to identify a tractable information-processing problem, but not yet task analysis (let alone empirical tests of the theory), this does not preclude Pietraszewski from evaluating whether the decades of existing data comport with his predictions. An accounting of the relevant evidence is, however, absent.

Here, I briefly review an alternative formsal account of social group discovery, in complement to Pietraszewski’s, which addresses these challenges. We adopt a computational model of latent structure learning to move beyond explicit category labels and dyadic similarity as the sole inputs to social group or coalition representations (Gershman & Cikara, 2020). By contrast to Pietraszewski’s model, latent structure learning (1) does not require conflict or ancillary cues (e.g., similarity on some feature), (2) does not require observation of overtly helpful or harmful behavior, which are relatively rare (Tooby & Cosmides, 2000), and (3) has already garnered empirical support.

We began by asking to what extent do people rely on similarity versus inferences of latent group structure, based on *observable behavior*, to guide their choice of allies? Note first that the behavior need not be help or harm, though it could be. Note also that the input here is not features (e.g., skin tone and language) which would be considered ancillary cues in Pietraszewski’s framework. A mere similarity account predicts that people substitute judgments of behavioral similarity to the self to identify allies (e.g., did this person vote for the same candidate I did in the last election?). This approach, however, neglects that people, including very young children, are sensitive to how well agents coordinate not just with themselves but with others in the environment indicating that people are predisposed to building representations of coordinated coalitions – or social structures – out in the world rather than just egocentric, dyadic similarities or interdependencies (see Cikara, 2021 for a recent review). Thus, an alternative hypothesis is that people’s inferences about coalition membership are improved by integrating information both about how agents relate to oneself as well as how they relate to one another.

By what process could people generate group representations on the basis of observing others’ non-coalitional behaviors? Via basic statistical learning algorithms. The normative solution to this inference problem is given by Bayes’ rule, which stipulates that the *posterior probability* over groupings given behaviors – P (grouping|behaviors) – is proportional to the product of the *likelihood* – P(behaviors|grouping) – and the *prior probability* P (grouping) (Gershman, Pouncy, & Gweon, 2017). Individuals who behave similarly will tend to be grouped together, but these groupings are dynamic and context dependent. As Pietraszewski (and we) argue, any model of group inference must be updated over time with more evidence and take into account the influence of all interagent relationships or similarities, not just mere similarity to oneself.

To adjudicate between mere similarity versus structure learning accounts we predicted that even when two agents' choices were equally similar to participants' own, participants' decisions would be influenced by the presence of a third agent who altered the coalitional structure (i.e., by creating a latent group that included the participant and only one of the first two agents). Importantly, a dyadic similarity account would predict that a third agent would have zero influence because the first two agents were equally similar to participants.

We tested this prediction in a series of behavioral experiments framed as learning about strangers' political issue positions (Lau, Pouncy, Gershman, & Cikara, 2018). In each trial, participants stated their position for or against a political issue, and then predicted the choices of three other agents on that same issue. After each prediction, they received feedback about that agent's actual choice. Finally, at the end of this learning phase, participants had to choose with which agent – A or B – they wanted to align themselves on a "mystery issue." Critically, agents A and B agreed with the participant an equal number of times during the learning phase, making them equally similar to the participant. Depending on the block, however, agent C either clustered with agent B *and* the participant, or only with agent B, excluding the participant.

As predicted by a latent structure learning account, participants were more likely to align themselves with agent B than A when C's placement created a cluster that put the participant in the same group as agent B (despite the fact that agents A and B were equally similar to the participant). Participants also judged agent B as more competent, moral, and likable than agent A when agent B clustered with the participant versus not. Perhaps most interesting, latent structures continued to exert an effect on ally-choice behavior even when we provided participants with explicit group labels that contradicted the latent structure (i.e., always put agent B in the explicit outgroup). In a companion functional magnetic resonance imaging (fMRI) experiment, the neural signals associated with latent structure representations further explained ally-choice behavior whereas interagent similarity-associated signal did not (Lau, Gershman, & Cikara, 2020).

In summary, and in line with Pietrasweski's challenge to the field, the latent structure learning framework moves away from ancillary similarity and category membership as sole inputs to group representation and inference, and retains the context sensitivity that is a major strength of Pietrasweski's account, but does not require conflict or observation of coalitional behaviors as inputs.

Financial support. This work was generously supported by the National Science Foundation (CAREER award, BCS-1653188).

Conflict of interest. The author claims no conflicts.

References

Cikara, M. (2021). Causes and consequences of coalitional cognition. *Advances in Experimental Social Psychology, 64*, 65–128.
Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science, 9,* 245–274.
Gershman, S. J., & Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science, 29,* 460–466.
Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science, 41,* 545–575.
Lau, T., Gershman, S.J., & Cikara, M. (2020). Social structure learning in human anterior insula. *eLife, 9*, e53162 (See "Spotlight" feature on this paper in Trends in Cognitive Sciences).
Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General, 147,* 1881–1891.
Tooby, J., & Cosmides, L. (2000). *Evolutionary psychology: Foundational papers.* MIT Press.

Private versus public: A dual model for resource-constrained conflict representations

Simon DeDeo[a,b]

[a]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15208, USA and [b]Santa Fe Institute, Santa Fe, NM 08541, USA sdedeo@andrew.cmu.edu; https://santafe.edu/~simon/ | doi:10.1017/S0140525X21001424, e102

Abstract

Pietraszewski's representation scheme is parsimonious and intuitive. However, internal mental representations may be subject to resource constraints that prefer more unusual systems such as sparse coding or compressed sensing. Pietraszewski's scheme may be most useful for understanding how agents communicate. Conflict may be driven in part by the complex interplay between parsimonious public representations and more resource-efficient internal ones.

Conflict is a matter of perception as well as action. Whether I attack, defend, or wait my turn, I must infer the new state of play in order to decide what to do next. If we want to understand this social feedback loop (Hobson, Mønster, & DeDeo, 2021), we have to get clear on what mental representations those inferences are actually about (DeDeo, 2017; Hobson & DeDeo, 2015). The limits and powers of those representations become guiding factors in how I act.

Pietraszewski's proposal is an atomic-compositional one. A mind represents the conflict in question by composing together units drawn from a small number of basic types. Tokens of these types, in turn, are triplets of "individuals" interacting in one of four patterns.

There is much to like about this idea. Its atoms are reminiscent of empirical results on conflict motifs (Shizuka & McDonald, 2012), for example. Most appealing, to my mind, is the sheer expressiveness of Pietraszewski's scheme. It seems very likely that, with enough time and patience, his proposed representation system can represent arbitrarily sophisticated conflicts.

But is this actually how the mind gets the job of representation done? Representations must answer to the mind's resource constraints – and a key lesson from the last decade is that the particular nature of those constraints can lead to unexpected cognitive effects (Shah & Oppenheimer, 2008).

In particular, constraints may well favor more profligate systems with a much larger set of basic symbols. Atomic-compositional systems are efficient in the number of basic types (the "dictionary"), but the size of that dictionary is not always the constraint that matters.

What may, in fact, matter more is not the number of *types*, but the number of *tokens* that any particular representation requires, a phenomenon known as sparse coding. The importance of sparse coding was first discovered in studies of the visual system, when

Olshausen and Field (1997) showed that low-level visual representations could be reconstructed under the assumption that the relevant constraint was not the number of neurons, but the energy required to fire them.

Sparse-coding dictionaries look very different from atomic-compositional ones. Among other things, they mix different levels: Basic units are neither individuals, nor coarse-grained entities, but combinations of both. If our internal representations make use of these sparse codes, then they would be expected to cut diagonally across Pietraszewski's taxonomy, potentially in complicated and hard-to-articulate ways.

Daniels, Krakauer, and Flack (2012) conducted the first empirical study of conflict to use sparse coding. The representations they extracted enabled high-fidelity reconstruction of real-world events, a basic adaptive goal for any representation system. As expected, the dictionary mixed levels, so that genetically related groups and unrelated individuals were simultaneously implicated in a single "symbol."

Such codes provide a counter-hypothesis to a priori atomic-compositional systems. They may also explain how participants can reason effectively across multiple scales; in our study on conflict in Wikipedia, for example, DeDeo (2016) found that antagonists compete against a background of hidden variables that characterize group-level activity. Turnover in these conflicts online may approach 70% per day (DeDeo, 2014) – meaning that after a few days the "same" conflict will have an entirely new collection of participants while maintaining recognizable patterns of strategic interaction.

Although sparse coding is a response to constraints on information processing, a related phenomenon, compressed sensing, is a potential response to constraints on information acquisition (Donoho, 2006). Compressed sensing schemes allow complex environments to be known "at a glance." When the representations being inferred have the correct properties, the number of observations required can be far fewer than the potential number of configurations. Compressed sensing may provide a (algorithmic) mechanism for the sensing of gestalt, and a second alternative to the atomic-compositional hypothesis

In either case, what makes the codes more efficient also makes them more difficult to communicate. A Wikipedia editor may make effective decisions by manipulating a sparse-coded representation, but the complexity and unusual nature of those codes may mean they are unable to express the basis of that decision to someone else.

This leads to a critical gap – and one that Pietraszewski's account may be able to fill. Conflict is, as the game theorists teach us, a fundamentally social matter. We don't simply fight with each other: we fight alongside each other. That places a premium on the ability to communicate and establish common knowledge about our beliefs.

The distinction between private and public representations is crucial. Mercier and Sperber's (2011) argumentative account, for example, distinguishes the basis on which we come to believe things, and the basis on which we argue for those beliefs to others. Because conflict requires communication among allies, it may be a terrific test case for the interaction of these two systems.

On the one hand, resource constraints lead to hard-to-articulate, but efficient, representations at the private, internal level. On the other hand, the need to communicate to others leads to a separate set of more easily grasped atomic-compositional representations such as Pietraszewski's. If I wish to draw someone into an alliance with me, I may well use language that maps quite closely onto Pietraszewski's four-term taxonomy.

Consider, for example, Machiavelli's analysis of conflict in The Prince. The general on horseback may have a head full of sparse codes, but Machiavelli's text is one Pietraszewskian story after another, a constellation of motifs of alliance, displacement, and betrayal. Indeed, Machiavelli's text often struggles to explain why sophisticated military strategists fail to make what are (to him) the most obvious moves – perhaps because those moves are obvious only when framed in an atomic-compositional framework, rather than in the internal representations that might be used by the participants themselves.

Whether or not Pietraszewski's framework matches the internal representations, a dual model leads us to ask what happens when allies and antagonists take the public representations seriously as a basis of action. The most interesting and adaptive features of conflict may arise precisely here, in the competition between the public and private, the parsimonious and efficient.

Financial support. This research was funded in part by Army Research Office Grant W911NF1710502 and by Jaan Tallinn via the Survival and Flourishing Fund.

Conflict of interest. None.

References

Daniels, B. C., Krakauer, D. C., & Flack, J. C. (2012). Sparse code of conflict in a primate society. *Proceedings of the National Academy of Sciences, 109*(35), 14259–14264.
DeDeo, S. (2014). Group minds and the case of Wikipedia. *Human Computation, 1*(1), 5–29.
DeDeo, S. (2016). Conflict and computation on Wikipedia: A finite-state machine analysis of editor interactions. *Future Internet, 8*(3), 31.
DeDeo, S. (2017). Major transitions in political order. In S. Walker, P. Davies, & G. Ellis (Eds.), *From matter to life: Information and causality* (pp. 393–428). Cambridge: Cambridge University Press. doi:10.1017/9781316584200.016.
Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory, 52*(4), 1289–1306.
Hobson, E. A., & DeDeo, S. (2015). Social feedback and the emergence of rank in animal society. *PLoS Computational Biology, 11*(9), e1004411.
Hobson, E. A., Mønster, D., & DeDeo, S. (2021). Aggression heuristics underlie animal dominance hierarchies and provide evidence of group-level social information. *Proceedings of the National Academy of Sciences, 118*(10), e2022912118.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74.
Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311–3325.
Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*(2), 207.
Shizuka, D., & McDonald, D. B. (2012). A social network perspective on measurements of dominance hierarchies. *Animal Behaviour, 83*(4), 925–934.

Are we there yet? Every computational theory needs a few black boxes, including theories about groups

Andrew W. Delton

Department of Political Science, College of Business, Stony Brook University, Stony Brook, NY 11794-4392, USA
andrew.delton@stonybrook.edu | https://www.andrewdelton.com | doi:10.1017/S0140525X21001217, e103

Abstract

Pietraszewski exemplifies the need for computational theory using group conflict; I complement this with an example of group cooperation. He criticizes past theories for having black boxes; I suggest his theory also has a black box – the concept of costs. He divides what mentally constitutes a group from mere ancillary attributes; I hazard that some of these attributes are essential.

Matter becomes mind when sculpted into a computer. To understand what our minds are doing, then, we need to understand the computations they perform. The target article by Pietraszewski illustrates the clarity won by computational theories, using the example of group conflict. I provide a complementary example using group cooperation.

As Pietraszewski points out, researchers often get caught up with the definitions that their methods make obvious. Consider research on group cooperation that uses the public goods game. The game reveals the tension between choosing what's good for your group and choosing what's good for you. Players start with a stake of real money that they can contribute to a collective fund. Contributions are multiplied; for instance, every $1 contributed becomes $2 in the fund. The fund is then divided equally among the players regardless of whether a player contributed. This simulates how people create greater benefits when they cooperate and how they can exploit one another. This is because contributing everything to the fund is collectively best, but each individual is even better off if they free ride on others' contributions. Cooperation won't last long if free riders persist, so people need to catch them and change their behavior or exclude them from the group.

When researchers use this game, they define a free rider as a player who (a) contributes less than others while (b) taking collective benefits. Using a computational lens, I've shown that neither (a) nor (b) is necessary or sufficient for the mind to categorize a person as a free rider; how the mind computes is more subtle (Delton, Cosmides, Guemo, Robertson, & Tooby, 2012; Delton, Nemirow, Robertson, Cimino, & Cosmides, 2013; Delton & Robertson, 2012; Delton & Sell, 2014). First, contributing less, in and of itself, does not matter; if someone contributes less by accident, they are still viewed as a cooperator. Second, a person can still be categorized as a free rider even when contributing equally – so long as they *wanted* to exploit the group. Third, a person need not actually take collective benefits to be a free rider; the mere possibility that they might later exploit the group is enough. Fourth, some cooperation functions as mutual aid. In relationships of mutual aid, people can take collective benefits without contributing, such as during illness or injury (Gurven, 2004; Sugiyama, 2004); treating these people as free riders would defeat the purpose of social insurance. The mind's definition of free rider does not neatly map onto definitions based on typical games. The computational theories Pietraszewski champions reveal what the mind is really up to.

Despite my enthusiasm for Pietraszewski's approach, perhaps he is being too hard on past ideas of what makes a group to the mind. I agree the containment metaphor fails. But it's less obvious that obligations or interdependence are poor theories when fleshed out beyond one-word summaries (Balliet, Tybur, & Van Lange, 2017). Pietraszewski argues that these ideas rely too much on intuition. His own theory, however, uses an intuitive idea: the concept of cost. Costs are not out there in the world; the mind must compute them. Daniel Kahneman and Amos Tversky famously pointed out the special psychology of costs and losses (e.g., Tversky & Kahneman, 1992). Even the most obvious cost – damaging the body – is not straightforward: If someone punctures your skin, isn't that a cost? Not when they're a doctor injecting a vaccine or palpitating a stopped heart. True, the needle and scalpel hurt, but they are unpleasant means to life-giving ends. One clue that something is a cost is that we get angry at the person who inflicts it (for a computational theory of anger and costs, see Sell et al., 2017). A patient doesn't get angry at the doctor who saved their life, even if at knifepoint. We lack a complete theory of how the mind defines costs (or anything else!). Given our ignorance, every computational theory leans on a few black boxes. Is a black box for costs much different than failing to have a complete theory of obligations or interdependence?

Finally, Pietraszewski wavers on whether conflict is required to make a group. Usually, he writes that he is only talking about groups-in-conflict; other times he seems to be referring to groups full stop. For instance, his lists "ancillary attributes" of groups – features that often go along with groups but do not *define* a group in the mind. Ancillary attributes include working together or sharing interests. I'm not sure such attributes are ancillary. What if there were groups without group conflict? Imagine a fantasy land where the *only* evolved function of groups is mutual aid. Even here, I would wager, the mind would still evolve the ability to see groups out in the world. If Jenny, Claire, and Raj help one another during illness – bringing food, paying bills, and so on – others would find it useful to know about this relationship. Although relationships of mutual aid have boundaries (there's that containment metaphor), their function is not to fight people outside the group but to help people inside it. Humans did evolve to compete in groups over status and resources; elsewhere I've argued, like Pietraszewski, that the mind evolved concepts that enable this competition (Cimino & Delton, 2010; Delton & Cimino, 2010; Delton, Kane, Petersen, Robertson, & Cosmides, 2021; Delton & Krasnow, 2017; Delton, Petersen, & Robertson, 2018). Human conflict often involves groups, groups often get into conflict, but groups and conflict are not identical.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Balliet, D., Tybur, J. M., & Van Lange, P. A. M. (2017). Functional interdependence theory: An evolutionary account of social situations. *Personality and Social Psychology Review, 21*(4), 361–388, https://doi.org/10.1177/1088868316657965.

Cimino, A., & Delton, A. W. (2010). On the perception of newcomers: Toward an evolved psychology of intergenerational coalitions. *Human Nature, 21*(2), 186–202.

Delton, A. W., & Cimino, A. (2010). Exploring the evolved concept of newcomer: Experimental tests of a cognitive model. *Evolutionary Psychology, 8*(2), 317–335.

Delton, A. W., Cosmides, L., Guemo, M., Robertson, T. E., & Tooby, J. (2012). The psychosemantics of free riding: Dissecting the architecture of a moral concept. *Journal of Personality and Social Psychology, 102*(6), 1252–1270.

Delton, A. W., Kane, J. V., Petersen, M. B., Robertson, T. E., & Cosmides, L. (2021). Partisans use emotions as social pressure: Feeling anger and gratitude at exiters and recruits in political groups. *Party Politics*, https://doi.org/10.1177/13540688211018796.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior, 38*(6), 734–743, https://doi.org/10.1016/j.evolhumbehav.2017.07.003.

Delton, A. W., Nemirow, J., Robertson, T. E., Cimino, A., & Cosmides, L. (2013). Merely opting out of a public good is moralized: An error management approach to cooperation. *Journal of Personality and Social Psychology, 105*(4), 621–638.

Delton, A. W., Petersen, M. B., & Robertson, T. E. (2018). Partisan goals, emotions, and political mobilization: The role of motivated reasoning in pressuring others to vote. *Journal of Politics, 80*(3), 890–902.
Delton, A. W., & Robertson, T. E. (2012). The social cognition of social foraging. *Evolution and Human Behavior, 33*, 715–725.
Delton, A. W., & Sell, A. (2014). The co-evolution of concepts and motivation. *Current Directions in Psychological Science, 23*(2), 115–120, https://doi.org/10.1177/0963721414521631.
Gurven, M. (2004). To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences, 27*(4), 543–583.
Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., … Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition, 168*, 110–128, https://doi.org/10.1016/j.cognition.2017.06.002.
Sugiyama, L. S. (2004). Illness, injury, and disability among Shiwiar forager-horticulturists: Implications of health-risk buffering for the evolution of human life history. *American Journal of Physical Anthropology, 123*(4), 371–389.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297–323.

Coalitionary psychology and group dynamics on social media

Jeff Deminchuk[a] and Sandeep Mishra[b]

[a]Department of Psychology, University of Regina, Regina, SK S4S 0A2, Canada and [b]Lang School of Business & Economics, University of Guelph, Guelph, ON N1G 2W1, Canada | [a]jeff.deminchuk@gmail.com | [b]sandeep.mishra@uoguelph.ca; sandeepmishra.ca |
doi:10.1017/S0140525X21001485, e104

Abstract

Pietraszewski's model allows understanding group dynamics through the lens of evolved coalitionary psychology. This framework is particularly relevant to understanding group dynamics on social media platforms, where coalitions based on salience of group identity are prominent and generate unique frictions. We offer testable hypotheses derived from the model that may help to shed light on social media behavior.

Pietraszewski's framework allows for the generation of novel hypotheses relevant to understanding group dynamics on social media platforms, where coalitions are prominent, problematic, and societally impactful. There is much debate on the precise nature of interactions of groups and individuals on social media, with often conflicting accounts (reviewed in Barberá, 2020). Some research suggests that social media facilitates "echo chambers" of amplified like-minded views; others have found that social media facilitates increased exposure to cross-cutting, diverse information (Cinelli, Morales, Galeazzi, Quattrociocchi, & Starmini, 2021; Nguyen & Vu, 2019). There is also debate regarding the specific effects of cross-cutting interactions, with evidence suggesting that such interactions can both reduce and exacerbate conflict (Paluck, Green, & Green, 2019). Understanding social media interactions through the lens of evolved coalitionary psychology may help to clarify and resolve many of these tensions.

Social media may amplify factors that exacerbate affective polarization, leading to intergroup conflict. Pietraszewski's model suggests that identity is the primary factor that scales up to define a group. The primacy of identity neatly accounts for the observation that social media (and, more generally, partisan media) tends to fuse political and social identity, as well as amplify identity awareness (Barberá, 2020; Hutchens, Hmielowski, & Beam, 2019; Iyengar, Lelkes, Levendusky, Malhotra, & Westwood, 2019; Iyengar, Sood, & Lelkes, 2012). Because of algorithms purpose-built to maximize "engagement," social media disproportionately elevates the voices of the most polarizing and hostile individuals, granting them great influence over others (Brady, Wills, Burkart, Jost, & Van Bavel, 2019; Iyengar et al., 2012; Shore, Baek, & Dellarocas, 2018). The "clickbait" driven business models of social media companies incentivizes threat-related, emotionally charged, and moralizing messages. Such messages appear to "hack" our evolved psychology, with such messages spreading more virally than neutral or positive messages (Berger & Milkman, 2012; Blaine & Boyer, 2018; Boyer, Firat, & van Leeuwen, 2015; Brady et al., 2019). People tend to dislike opposing elites more than the average outgroup member, while simultaneously conflating their own views with their group's views. Consequently, repeated exposure to inflammatory messaging may be a salient factor contributing to increased affective polarization (Druckman & Levendusky, 2019). More generally, competition and perceived disadvantage may contribute to the elevation of risky, antisocial, and/or uncooperative behaviors both on and off social media (Mishra, Barclay, & Sparks, 2017).

For individual coalition members, social media allows for amplified social signaling. Actors can gain outsized reputational benefits by broadcasting to an audience far larger than any experienced in ancestral environments, thereby receiving disproportionate social reinforcement via "likes," "reactions," or other socially engineered forms of operant conditioning (Dos Santos & Rankin, 2010; Petersen, 2015; Tooby & Cosmides, 2010). Although signaling is not necessarily related to conflict, the factors described above, along with a significantly diminished possibility of incurring retaliatory costs, may create an environment where rewards for hostile intergroup signals are artificially inflated while costs are artificially suppressed.

Pietraszewski's model suggests a set of falsifiable hypotheses regarding the causes of social media amplification of coalitional conflict. News organizations and other actors with financial incentives tied to social media engagement disproportionately broadcast moral-emotive or threat-related messages to their followers. These signals function to mentally co-register (coordinate) followers, increase emotional engagement, amplify the salience of group identity, and facilitate the adoption of one of the four group-roles (generalization, alliance, defense, and displacement) that the model identifies (Barberá, 2020; Petersen, 2015; Tooby & Cosmides, 2010). The average group member is more inclined to act (consistent with one's evoked role, as predicted by the model) propagating conflict. From the outgroup's perspective, attacks increase threat perceptions, placing outgroup members in a defensive posture. This process makes the formation of a defensive coalition more likely, with an associated greater likelihood of preemptive attacks (Karlsen, Steen-Johnsen, Wollebæk, & Enjolras, 2017). This cycle becomes self-perpetuating. This hypothesized cascade can readily explain why cross-cutting information, typically associated with a reduction in conflict/polarization, has been recently found to have the opposite effect (Barberá, 2020).

It would be illuminating to test the above cascade by examining social media behavior of different coalitions following a polarizing event that sets groups in opposition (a recent example being the murder of George Floyd). Such high-profile events offer salient opportunities for influential group members to coordinate a punitive coalition with moral-emotive or threat-related messages (Tooby & Cosmides, 2010). Such circumstances could also be used to investigate dynamics between groups that do not typically associate. For example, during a recent violent flare up in conflict between Palestine and Israel, leaders and news organizations tied to the Black Lives Matter movement shared posts in support of Palestinian causes with higher frequency, whereas those associated with a group often acting in opposition to Black Lives Matter activists – Donald Trump supporters – signaled pro-Israeli sentiments with higher frequency. Consistent with Pietraszewski's model, we would expect to observe a clear increase in willingness to impose costs on the opposing groups in support of their allies mediated by an increased identification with the allied group.

Another potential area of exploration relevant to Pietraszewski's model is the effect of removing prominent individuals from platforms. The model suggests that removing a prominent node may serve as an effective form of coalitional warfare by decreasing the opposing group's co-registration and eroding their group identity. However, target group members may see this act as an attack in and of itself, which would have the effect of strengthening group identity, allowing for the more expedient formation of a defensive coalition (Petersen, 2015). Both effects may also occur with group identity being strengthened in the short term based on the perceived threat, but weakened over the long term as a unifying voice is lost.

Pietraszewski's model specifically, and the lens of evolved coalitionary psychology more generally, may help to clarify group dynamics and resultant conflict on social media. These frameworks may help to clarify much of the uncertainty in understanding of causality that exists in the social media landscape, helping to engineer online environments that may help de-escalate conflict, rather than inflame it.

Financial support. This work was supported by an Insight Grant to SM from the Social Sciences and Humanities Research Council of Canada, no. 430718.

Conflict of interest. JD and SM have no conflicts of interest to declare.

References

Barberá, P. (2020). Social media, echo chambers, and political polarization. In Persily, N., & Tucker, J. (Eds.), *Social media and democracy: The state of the field* (pp. 34–55). Cambridge University Press. https://doi.org/10.1017/9781108890960.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192–205. https://doi.org/10.1509/jmr.10.0353.

Blaine, T., & Boyer, P. (2018). Origins of sinister rumors: A preference for threat-related material in supply and demand of information. *Evolution and Human Behavior, 39*(1), 67–75. https://doi.org/10.1016/j.evolhumbehav.2017.10.001.

Boyer, P., Firat, R., & van Leeuwen, F. (2015). Safety, threat, and stress in intergroup relations: A coalitional index model. *Perspectives on Psychological Science 10*(4), 434–450. https://doi.org/10.1177/1745691615583133.

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General, 148*(10), 1802–1813. https://doi.org/10.1037/xge0000532.

Cinelli, M., Morales, F. G. D., Galeazzi, A., Quattrociocchi, W., & Starmini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences, 118*(9), e2023301118. https://www.doi.org/10.1073/pnas.2023301118.

Dos Santos, M., & Rankin, D. J. (2010). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences 278*, 371–377. https://doi.org/10.1098/rspb.2010.1275.

Druckman, J. N., & Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly, 83*(1), 114–122. https://www.doi.org/10.1093/poq/nfz003.

Hutchens, M. J., Hmielowski, J. D., & Beam, M. A. (2019). Reinforcing spirals of political discussion and affective polarization. *Communication Monographs, 86,* 357–376. https://www.doi.org/10.1080/03637751.2019.1575255.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science, 22,* 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly, 76,* 405–431 https://www.doi.org/10.1093/POQ/NFS038.

Karlsen, R., Steen-Johnsen, K., Wollebæk, D., & Enjolras, B. (2017). Echo chamber and trench warfare dynamics in online debates. *European Journal of Communication, 33* (3), 257–273. https://doi.org/10.1177/0267323117695734.

Mishra, S., Barclay, P., & Sparks, A. (2017). The relative state model: Integrating need-based and ability-based pathways to risk-taking. *Personality and Social Psychology Review, 21,* 176–198.

Nguyen, A., & Vu, H. T. (2019). Testing popular news discourse on the "echo chamber" effect: Does political polarisation occur among those relying on social media as their primary politics news source? *First Monday, 24*(6). https://doi.org/10.5210/fm.v24i6.9632.

Paluck, E., Green, S., & Green, D. (2019). The contact hypothesis re-evaluated. *Behavioural Public Policy, 3*(2), 129–158. https://www.doi.org/10.1017/bpp.2018.25.

Petersen, M. B. (2015). Evolutionary political psychology. In Buss, D. M. (Ed.) *The handbook of evolutionary psychology* (2nd ed., pp. 1084–1102). Wiley.

Shore, J., Baek, J., & Dellarocas, C. (2018). Network structure and patterns of information diversity. *MIS Quarterly, 42*(3), 849–872. https://www.doi.org/10.25300/MISQ/2018/14558.

Tooby, J., & Cosmides, L. (2010). Groups in mind: Coalitional psychology and the roots of war and morality. In Høgh-Olesen, H. (Ed.), *Human morality and sociality: Evolutionary and comparative perspectives* (pp. 191–234). Palgrave Macmillan.

Psychological and actual group formation: Conflict is neither necessary nor sufficient

Julia Elad-Strenger[a] and Thomas Kessler[b]

[a]Department of Political Studies, Bar-Ilan University, Ramat Gan 5290002, Israel and [b]Department of Psychology, Friedrich-Schiller-University Jena, 07743 Jena, Germany Eladstj@biu.ac.il | thomas.kessler@uni-jena.de | https://www.juliaeladstrenger.com/ | http://www.sozialpsychologie.uni-jena.de | doi:10.1017/S0140525X21001321, e105

Abstract

Conflict is neither necessary nor sufficient for the existence of groups. First, the existence of mutually *supporting*, rather than antagonistic, interactants is sufficient to constitute a "social group." Second, conflict does not necessarily mark group boundaries but can also exist within an ingroup. Third, psychological representations of social groups do not only trace, but also perpetuate the existence of groups.

It is clearly difficult to define a "social group," as there are several aspects that must be explained for a comprehensive understanding of the concept. Psychological definitions converge on the categorization of self and others into groups, which produces perceptions of groups or group behaviour (e.g., mutual support).

Sociology and anthropology characterize groups by shared norms, standards, and institutions regulating the behaviour of group members.

Adding an evolutionary perspective to the existing literature, the present article proposes a computational theory of social groups that identifies four triadic primitives, defining specific group-constitutive roles, which represent "social groups" in the human mind. These primitives include three actors and the interaction between them. In all these triads, two of the interactions are assumed to be negative (e.g., attack and threat), whereas one is assumed to be positive (e.g., support). In this commentary, we focus on the author's suggestion that the negative relations are the group-constitutive factor within these triads.

Our first argument is that conflict is not a necessary component (or marker) for the constitution of groups. The question whether conflict is indeed a necessary component of groups can be traced back to the debate between the Darwinian and Kropotkinian perspectives (Todes, 1987). Although Darwin focused on the "law of mutual struggle" as the driving force in evolution, Kropotkin believed that this law is complemented by a "law of mutual aid" (e.g., Skyrms, 2014). In line with Kropotkin, we argue that not the "attack" relations but the "mutually-supporting" relations are the actual primitives that are traced by mental representations of a "social group" (Brown, 1988; Kessler & Cohrs, 2008; Tajfel & Turner, 1979; Turner, 1982). The existence of two or more interactants that are in mutually supporting relations (compared to no relations or antagonistic relations) would be sufficient to constitute a group and represent it.

According to this perspective, group members would reliably support each other when one of them (or all as a collective) face a challenge, even if such a challenge is not posed by an antagonistic individual or group (Elad-Strenger, 2016). As an example, team members working to achieve a common goal constitute a group that is detectable and represented as such, even without being attacked from the outside. Thus, although we agree that the notion of conflict with outsiders may enhance mutual support and commitment within the group (Elad-Strenger, 2013; Elad-Strenger & Shahar, 2017; Sherif, 1966), we argue that conflict is not necessary for the psychological representation and the actual existence of a group. Group boundaries can just as well be located where the mutual support ends and neutral (e.g., indifference) or negative relations begin (e.g., attack).

Our second argument is that the existence of conflict within a triad does not necessarily create or signal the existence of antagonistic groups (those who are inside the group and those who are outside), but may simply represent an internal dynamic within an ingroup, which may even strengthen the agents' belongingness to the group. An example for such a dynamic is ingroup deviance and group members' response to the deviance. The specific norms of the group define what is considered an aggressive or deviant act towards the group (Ben-Shitrit, Elad-Strenger, & Hirsch-Hoefler, 2021; Elad-Strenger, Hall, Hobfoll, & Canetti, 2021). Accordingly, some deviances will be met with the exclusion of the deviant from the group, whereas many others will trigger attempts to reform the deviant (Kessler & Cohrs, 2008). Ingroup deviants are remembered better (Hechler, Neyer, & Kessler, 2016) and punished harsher (e.g., Marques and Yzerbyt, 1988) than outgroup deviants, precisely because the deviants and respondents are members of the same group. Thus, it is not only that groups can exist despite occasional internal deviance or conflict (Elad-Strenger, Fireman, Schiller, Besser, & Shahar, 2013; Elad-Strenger, Halperin, & Saguy, 2019), but also that the existence of deviance may even strengthen ingroup members' identification with it (Pinto, Marques, Levine, & Abrams, 2010).

What, then, defines a "group" as such, despite internal conflicts? Our third argument is that a mental representation of a group does not only reflect its existence, but also produces and perpetuates it (Turner & Giles, 1981), by means of a "self-fulfilling prophecy." The self-categorization of two agents as ingroup members tends to produce coordination and cooperation between these agents, and antagonistic behaviours towards agents who are categorized as outsiders, thus forming a group via behavioural confirmation (Sassenberg, Kessler, & Mummendey, 2003; Snyder & Swann, 1978; Tajfel, Billig, Bundy, & Flament, 1971). This mental representation of a group can even define ingroup and outgroup members despite the behavioural primitives would point the other way. For example, when police officers are attacked, they defend one another (signalling the existence of a group) but also defend (and are defended by) their police dogs. Nonetheless, when talking about the police, people rarely think about the dogs as belonging to this group, despite defending policewomen against attack, simply because they are not included in the mental representation of the group. In short, the mental representation of a group can determine the behaviours of agents and thus the existence of groups, as well as the extent to which these behaviours are interpreted as signalling the existence of a group. Considering this role of mental representations (i.e., cognitive categorization) in group constitution, attack or defence in response to conflict are, therefore, also not sufficient to constitute (or mark) groups.

To summarize, we argue that conflict within triadic relations is neither necessary nor sufficient to constitute groups. Rather, we suggest that mutual support is the relevant ecological invariance to trace the existence of groups. In addition, ingroup norms define which behaviours categorize the actors as ingroup or outgroup members. Finally, the mental representations of a social group determine, at least partially, the behavioural primitives and their interpretation. Therefore, we propose that the combination of different approaches to social groups (psychological: cognitive categorization; sociological and anthropological: shared norms; evolutionary: ecological invariances) can paint a more dynamic picture of groups than is represented in the computational theory of groups.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Ben-Shitrit, L., Elad-Strenger, J., & Hirsch-Hoefler, S. (2021). "Pinkwashing" the radical-right: Gender and the mainstreaming of radical-right policies and actions. *European Journal of Political Research, 61,* 86–110. https://doi.org/10.1111/1475-6765.12442.

Brown, R. (1988). *Group processes: Dynamics within and between groups* (2nd ed.). Blackwell.

Elad-Strenger, J. (2013). Changing minds: A psychodynamic interpretation of Kuhnian paradigm change. *Review of General Psychology, 17,* 40–52.

Elad-Strenger, J. (2016). Activism as a heroic quest for symbolic immortality: An existential perspective on collective action. *Journal of Social and Political Psychology, 4,* 44–65.

Elad-Strenger, J., Fireman, Z., Schiller, M., Besser, A., & Shahar, G. (2013). Risk-resilience dynamics of ideological factors in distress after the evacuation from Gush Katif. *International Journal of Stress Management, 20*, 57–75.

Elad-Strenger, J., Hall, B. J., Hobfoll, S., & Canetti, D. (2021). Explaining public support for violence against politicians during conflict: Evidence from a panel study in Israel. *Journal of Peace Research, 58,* 417–432.

Elad-Strenger, J., Halperin, E., & Saguy, T. (2019). Facilitating hope among the hopeless: The role of ideology and moral content in shaping reactions to internal criticism in the context of intractable conflict. *Social Science Quarterly, 100,* 2425–2444.
Elad-Strenger, J., & Shahar, G. (2017). Revisiting the effects of societal threat perceptions on conflict-related political positions: A three-wave study. *Journal of Conflict Resolution, 62,* 1753–1783.
Hechler, S., Neyer, F., & Kessler, T. (2016). The infamos among us: Enhanced reputational memory for uncooperative ingroup members. *Cognition, 157,* 1–13.
Kessler, T., & Cohrs, J. C. (2008). The evolution of authoritarian processes: How to commit group members to group norms. *Group Dynamics Theory, Research, and Practice, 12,* 73–84.
Marques, J. M., & Yzerbyt, V. Y. (1988). The black sheep effect: Judgmental extremity towards ingroup members in inter- and intragroup situations. *European Journal of Social Psychology, 18,* 287–292.
Pinto, I. R., Marques, J. M., Levine, J. M., & Abrams, D. (2010). Membership status and subjective group dynamics: Who triggers the black sheep effect? *Journal of Personality and Social Psychology, 99,* 107–119.
Sassenberg, K., Kessler, T., & Mummendey, A. (2003). Less negative = more positive? Social discrimination as avoidance and approach. *Journal of Experimental Social Psychology, 39,* 48–58.
Sherif, M. (1966). *In common predicament. Social psychology of intergroup conflict.* Houghton & Mifflin.
Skyrms, B. (2014). *Evolution of the social contract* (2nd ed.). Cambridge.
Snyder, M., & Swann, W. B. (1978). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology, 14,* 148–162.
Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology, 1*(2), 149–178.
Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In S. Worchel, & G. Austin (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks.
Todes, D. (1987). Darwin's Malthusian metaphor and Russian evolutionary thought, 1859–1917. *Isis, 78,* 537–551.
Turner, J. C. (1982). Towards a cognitive redefinition of the social group. In H. Tajfel (Ed.), *Social identity and intergroup relations* (pp. 15–40). Cambridge University Press.
Turner, J. C., & Giles, H. (1981). Introduction: The social psychology of intergroup behaviour. In J. C. Turner, & H. Giles (Eds.), *Intergroup behaviour* (pp. 1–32). Basil Blackwell.

Interacting with others while reacting to the environment

Ilan Fischer[a], Simon A. Levin[b], Daniel I. Rubenstein[b], Shacked Avrashi[a], Lior Givon[a], and Tomer Oz[a]

[a]Department of Psychology, University of Haifa, Mount Carmel Haifa 3498838 Israel and [b]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544-2016, USA | ifischer@psy.haifa.ac.il | slevin@princeton.edu | dir@princeton.edu | savrashi@campus.haifa.ac.il | lgivon02@campus.haifa.ac.il | toz04@campus.haifa.ac.il | doi:10.1017/S0140525X21001291, e106

Abstract

Here, we revise Pietraszewski's model of groups by assigning participant pairs with two triplets, denoting: (1) the type of game that models the interaction, (2) its critical switching point between alternatives (i.e., the game's similarity threshold), and (3) the perception of strategic similarity with the opponent. These triplets provide a set of primitives that accounts for individuals' strategic motivations and observed behaviors.

Aiming to define the primitives that comprise the notion of a group-in-conflict, Pietraszewski proposes to assign agents to one of four triadic interaction types (*Generalization, Alliance, Displacement,* and *Defense*), each involving three agents, two of them attacking another agent.

But, although an attack is sometimes an inevitable consequence of a conflict, cooperation is not less critical. Cooperation enables to allocate resources, pool skills and forces, and build coalitions. Moreover, although attacking and cooperating are apparent behaviors, or consequences, they result from different motivations and perceptions, which may exist long before they develop into observable actions. Specifically, groups are comprised of *individuals* that are simultaneously: (1) motivated by the aspiration to maximize subjectively perceived outcomes, (2) take into account the motivations and expected actions of in- and outgroup members, and (3) apply cognitive skills, such as comparing, interpreting, and planning. Hence, to define the primitives of group behavior, agents should not only be represented by apparent states, but also by their underlying motivations and perceptions.

It is important to note that group boundaries may mean different things. When studying existing groups, defined by social, geographic, or ethnic characteristics, these predefined features also define the groups' boundaries. But, when studying group dynamics and evolution, group boundaries develop as emergent patterns (Fischer et al., 2013; McIntosh, 2010).

Regardless of whether group structures are studied as a cause or as an effect, they involve at least two fundamental aspects: (1) the motivations induced by the expectations for material and social outcomes, and (2) the expectations for strategic conduct of friends and foes. As noted by Pietraszewski, a fixed set of payoffs, such as those comprising the prisoner's dilemma (PD) (Axelrod, 1984; Flood & Dresher, 1952; Rapoport & Chammah, 1965) or the chicken game (Rapoport & Chammah, 1966) may not always be suitable for the study of specific groups. But this does not imply that a comprehensive set of games is not a proper tool for the study of groups and their actions. On the contrary, over seven decades of *theoretical* and *behavioral* game theory research have provided studies of strategic insights and taxonomies of various interaction types (e.g., Rapoport, 1960; Rapoport & Guyer, 1966; von Neumann & Morgenstern, 2007).

To follow on Pietraszewski's notion and identify fundamental group interaction types, we propose to characterize each interaction by *two* triplets, each associated with the *perceptions* of one party. Each triplet comprises: the *type of game*, its *defining characteristics*, and the *strategic perception of the opponent*. To provide concise yet strategically detailed game descriptions, we apply Rapoport & Guyer's (1966) taxonomy of games that reduces social interactions into a set of 78 two-by-two *rank-ordered* payoff matrices, each exhibiting a set of strategic properties. For the purpose of associating game structures with both interpersonal and intergroup perceptions of the interacting partner (either a friend or a foe), we apply the theory of subjective expected relative similarity (SERS, Fischer, 2009, 2012) that provides both a normative solution and a descriptive, empirically validated, model. Unlike Pietraszewski, we do not assume that the applied model embodies a representation of the mind, but expect it to provide testable and valid hypotheses.

For example, consider two players interacting in a PD game defined by four payoffs: T, R, P, and S (Fig. 1). Each player, consciously or unconsciously, assigns the probability p_s to the prospects of the opponent choosing a similar alternative (and the

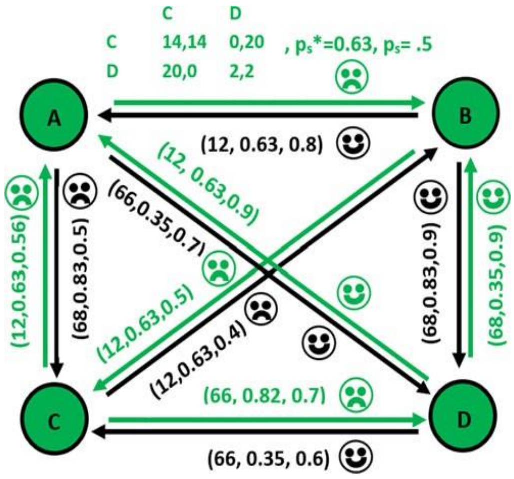


Figure 1 (Fischer et al.). Fundamental bidirectional group interaction types, reflecting the motivations and perceptions within pairs of interacting individuals (A, B, C, and D), each denoted by a unique triplet of elements that comprise: The number of the game the individual assumes he/she is playing in accord to Rapoport and Guyer (1966) taxonomy of two-by-two games (except for the upper pair which shows the entire PD game structure), the similarity threshold of the game (p_s*), and the perceived strategic similarity with the opponent (p_s). In addition, smiley (whenever $p_s > p_s$*) and frowny (whenever $p_s < p_s$*) faces denote observed or expected actions derived from the underlying motivations and perceptions. For games 12 (PD) and 66 (chicken) a smiley and a frowny face denote cooperation and defection, for game 68, which is a coordination game, a smiley and a frowny face denote the strategy that maximizes expected payoffs under sufficiently high similarity with the opponent and the strategy that maximize expected payoffs under non-sufficiently high similarity with the opponent.

complementary probability $1 - p_s$ to the prospects of the opponent choosing a dissimilar alternative). Comparing the expected values (EV) for the choices of *cooperation* ($Rp_s + S(1 - p_s)$) and *defection* ($Pp_s + T(1 - p_s)$) allows choosing the strategy that provides the higher EV. SERS allows computing a switching point between which of the two alternative strategies of the game should be favored (if such a point exists in the game), namely the *similarity threshold of the game*, denoted by p_s*. For example, considering the PD game and assuming EV(cooperation) = EV(defection), we obtain $p_s{*} = (T - S)/(T - S + R - P)$. By presenting and comparing p_s* with p_s, as perceived by each individual, we define all fundamental interaction types, comprising also the four types proposed by Pietraszewski. Importantly, the perception of strategic similarity with the opponent may relate to a specific individual or to a specific group, depending on what is modeled or empirically estimated.

Merging Rapoport & Guyer's (1966) taxonomy of games with SERS, we revise Pietraszewski's proposed model, by assigning each pair of participants with two triplets, one per participant. The triplets comprise: (1) the type of game that models the perceived interaction, denoted by a number from Rapoport & Guyer's (1966) taxonomy, which provides many established game theoretic insights, (2) the similarity threshold of the game, p_s* (derived from the exact and continuous payoffs perceived by the player), and (3) the perception of strategic similarity with the opponent, p_s. Figure 1 represents all, cooperative and hostile, symmetric, and asymmetric, fundamental interaction types. Among others it also shows the four types proposed by Pietraszewski, all exhibited by the actions of agents A, B, and C. The theoretic example shows agents that assume they are playing one of three games, either PD, chicken, or a coordination game. Sometimes both agents assume they are playing the same game and sometimes they assume they are playing a different game, as reflected by the first element of their assigned triplets. Even when agents play the same game, they may still differ in respect to the exact payoffs, which give rise to different p_s* values, denoted by the second element of the triplets. Agents may also differ in their perception of strategic similarity with the opponent, p_s, denoted by the third element of the triplets. Hence, even players that assume they play the same game with an identical p_s*, may still differ in their perceptions of strategic similarity with the opponent, and choose different actions.

Finally, we point to the possibility of further reducing triplets into pairs of p_s* and p_s values, which explain and predict all cooperative and competitive actions of group members.

Financial support. This research was supported by ISF grant No. 640/15 Forecasting Inter-group conflict potentials, The Army Research Office Grant W911NF-18-1-0325, and Princeton University.

Conflict of interest. The authors declare having no conflict of interests.

References

Axelrod, R. M. (1984). *The evolution of cooperation.* Basic Books.
Fischer, I. (2009). Friend or foe: Subjective expected relative similarity as a determinant of cooperation. *Journal of Experimental Psychology: General, 138*(3), 341–350. https://doi.org/10.1037/a0016073.
Fischer, I. (2012). Similarity or reciprocity? On the determinants of cooperation in similarity-sensitive games. *Psychological Inquiry, 23*(1), 48–54. https://doi.org/10.1080/1047840X.2012.658004.
Fischer, I., Frid, A., Goerg, S. J., Levin, S. A., Rubenstein, D. I., & Selten, R. (2013). Fusing enacted and expected mimicry generates a winning strategy that promotes the evolution of cooperation. *Proceedings of the National Academy of Sciences of the USA, 110,* 10229–10233. https://doi.org/10.1073/pnas.1308221110.
Flood, M., & Dresher, M. (1952). *Some experimental games. Research memorandum RM-789.* Rand.
McIntosh, H. V. (2010). Conway's Life. In A. Adamatzky (Ed.), *Game of life cellular automata* (pp. 17–33). Springer. https://doi.org/10.1007/978-1-84996-217-9.
Rapoport, A. (1960). *Fights, games, and debates. In Fights, games, and debates.* University of Michigan Press. https://doi.org/10.3998/mpub.9022.
Rapoport, A., & Chammah, A. M. (1965). *Prisoner's dilemma: A study in conflict and cooperation.* University of Michigan Press.
Rapoport, A., & Chammah, A. M. (1966). The game of chicken. *American Behavioral Scientist, 10*(3), 10–28. https://doi.org/10.1177/000276426601000303.
Rapoport, A., & Guyer, M. (1966). A taxonomy of 2 × 2 games. *General Systems, 11,* 203–214.
von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior* (60th Anniversary Commemorative Edition, Vol. 9781400829). Princeton University Press. https://doi.org/10.1515/9781400829460.

Internal versus external group conflicts

Agner Fog

Engineering Technology, Technical University of Denmark, Campus Ballerup, Lautrupvang 15, 2750 Ballerup, Denmark agner@agner.org | https://www.agner.org | doi:10.1017/S0140525X21001230, e107

Abstract

A group in intergroup conflict needs to overcome the collective action problem in order to defend itself against an external enemy. This leads to increasing complexity that cannot be adequately covered by just scaling up the model of intragroup conflicts. Research on cultural evolution and evolutionary psychology shows that external conflict has profound effects on group organization.

Pietraszewski should be complimented for trying to obtain order and system in the chaotic concept of group dynamics.

I think the model has a weakness in the scaling up of the triadic primitives to interactions between groups. The difference between intragroup conflict and intergroup conflict is not just a matter of scale. The whole social organization is often radically different between groups dominated by intragroup conflict and groups dominated by intergroup conflict (Fog, 2017).

The article mentions that subgroups opposing each other in a civil war may unite against an external common enemy. But the model does not adequately cover the complexity that can be observed in groups facing external enemies. Research on cultural evolution and evolutionary psychology shows that violent conflicts between human groups can lead to an arms race that drives a development toward increasing complexity. The group that is best at overcoming the collective action problem and motivate its members to risk their lives in a fight for their group is likely to win territory from a less organized enemy group. This requires a strong hierarchy, discipline, and a strong leader who can punish free riders, reward brave warriors, and organize the fighting. History shows that this mechanism can drive a cultural evolution toward ever-growing political units and increasing complexity. The strongest groups need an increasing division of labor with leaders, bureaucrats, police, judges, and of course soldiers, weapon producers, and food producers. They also tend to develop a political organization, culture, religion, and ideology that supports and motivates the collective fighting.

Groups with only internal conflicts and no external enemies tend to develop in a very different direction. Members of such groups will not tolerate a strong hierarchy with a kleptocratic tyrant at the top. Instead, we can observe that groups in peaceful surroundings tend to develop in the direction of egalitarian, tolerant, and democratic cultures (Fog, 2017).

If we want to build a theoretical model that describes groups in external conflict then we need to include the complexities of the collective action problem and the mechanisms necessary for making people sacrifice themselves for their group. The theory presented by Pietraszewski can be useful for game-theoretical models of groups with only intragroup conflicts, but it cannot adequately capture the drive toward increasing complexity that we have observed in groups dominated by intergroup conflicts.

Financial support. This research received no grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

Reference

Fog, A. (2017). *Warlike and peaceful societies: The interaction of genes and culture.* Open Book. doi: 10.11647/OBP.0128

The labelled container: Conceptual development of social group representations

Rebekah A. Gelpi[a], Suraiya Allidina[a], Daniel Hoyer[b] and William A. Cunningham[a]

[a]Department of Psychology, University of Toronto, Toronto, ON M5S 3G3 Canada and [b]Evolution Institute & Center for Preparatory and Liberal Studies, George Brown College, Toronto, ON M5A 3W8, Canada
rebekah.gelpi@mail.utoronto.ca | suraiya.allidina@mail.utoronto.ca | dhoyer@evolution-institute.org | wil.cunningham@utoronto.ca |
doi:10.1017/S0140525X21001412, e108

Abstract

Pietraszewski contends that group representations that rely on a “containment metaphor” fail to adequately capture phenomena of group dynamics such as shifts in allegiances. We argue, in contrast, that social categories allow for computationally efficient, richly structured, and flexible group representations that explain some of the most intriguing aspects of social group behaviour.

Pietraszewski in the target article offers a bottom-up approach to understanding the problem of group representation, describing how these representations could be constructed *ex nihilo* in the context of conflict by identifying patterns of interactions within triads. Although this account provides a straightforward computational account of mental representations of intergroup conflict, it neglects important top-down influences such as induction and generalization. We suggest that these social group categorization processes may better account for core features of modern human group living, such as flexible, dynamic social identities, and generalized trust in strangers in massive environments, than an approach that relies only on constructing group representations out of an event framework.

As bottom-up processing alone is unlikely to be realizable at scale given humans’ limited cognitive resources, top-down processes likely must play some role in developing group representations. One such process is inductive reasoning, which allows us to efficiently learn complex concept and category knowledge from relatively sparse data (Kemp, Perfors, & Tenenbaum, 2007; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Although categories can be quickly built from experience, we also receive rich, highly structured information about these categories in the form of generic language and category labels, quickly scaffolding our learning about abstract category structures, both non-social (Butler & Markman, 2014; Taverna, Padilla, Baiocchi, & Peralta, 2021) and social (Baron, Dunham, Banaji, & Carey, 2014; Gelman, Ware, & Kleinberg, 2010; Roberts, Ho, & Gelman, 2017b). By representing groups as abstract, symbolic categories that capture statistical regularities and make probabilistic predictions about how members are likely to appear and act, we gain several insights that are not captured by a model which constructs group representations out of event frameworks alone. Relying on category and concept learning not only allows us to use a highly tractable, domain-general strategy to learn about and represent complex social groups, but it also readily expands the predictive capacity of group assignment beyond conflict and reciprocity, to resolving highly abstract cooperation and coordination problems that complex societies must solve.

Beyond simply giving "rules for assignment" based on statistical regularities in shared features and behaviours, categories allow us to make predictions about how category members are likely to behave, and even to develop prescriptive norms about how they *ought* to behave (Foster-Hanson & Rhodes, 2019; Roberts, Gelman, & Ho, 2017a), including group members' moral obligations to the group (Chalik & Rhodes, 2018, 2020). In the context of a conflict, knowledge about agents' social categories can allow observers to make inferences about what kinds of behaviours or group-constitutive roles are more or less likely for an agent to take; rather than (or in addition to) the roles determining the inferences about groups, the prior expectations around group memberships determine the inference of roles.

Indeed, these processes may be critical in the formation and representation of groups where most members are strangers, as members of modern social groups such as cities and countries are able to recognize only a small proportion of the group's total population with fine individual detail (Dunbar, 2010), and must regularly interact with group members about which one has no information beyond knowledge of perceptually or contextually inferred social categories (such as race, gender, or nationality). Relying on group-based norms and expectations allows us to make broad inferences about strangers that facilitate social prediction without needing to represent these individuals or interactions in a deep way. The cultural transmission of norms regarding social roles and division of labour (e.g., Lew-Levy, Lavi, Reckin, Cristóbal-Azkarate, & Ellis-Davies, 2018) can also easily be accommodated within a framework of social category learning and, we argue, is in fact essential to explain how individuals access and internalize these norms within large, complex societies (Gavrilets & Richerson, 2017).

In contrast to an event model, categorization processes are likely more involved in the social learning processes that underlie cultural learning. Rather than being learned from the bottom-up, many group identities have been developed over long histories of cooperation, affiliation, or conflict. Although processes of social categorization are efficient for making sense of the social world, they come at a cost; when applied to these "prepackaged" groups, the cognitive tools of induction, categorization, and pedagogical reasoning that allow us to swiftly bootstrap sophisticated social category structures can lead to overgeneralization and stereotyping (Macrae, Milne, & Bodenhausen, 1994).

Interestingly, despite Pietraszewski's claim that categorizations are rigid, our social categories are not only swiftly and efficiently learned, but also highly flexible, dynamic, and complex in structure, with people holding multiple social identities which can overlap to varying degrees or even be contradictory (Roccas & Brewer, 2002). Social identities can rapidly shift in accordance with their context or prominence, and these shifts can, in turn, shape attention, modulating our memory for social categories (Van Bavel & Cunningham, 2012; Van Bavel, Packer, & Cunningham, 2011) as well as amygdala activation (Cunningham, Van Bavel, & Johnsen, 2008). When group identities are multiply nested or produce conflicting loyalties, contextual activation or prominence of a group identity could lead to using different social category structures to make behavioural decisions as well as predictions about likely outcomes in complex, real-world contexts.

We applaud Pietraszewski's goal of formalizing the process of social group representation and grounding our understanding of groups in this psychological process. Although we agree that triadic event information can motivate people's reasoning about some – particularly ad hoc – groupings, we believe that representing groups as categories provides unique insight into phenomena such as the transmission of cultural knowledge and generalized trust while allowing for symbolic, abstract group identities that are nevertheless flexible and situationally dependent. These features, we argue, may help us better understand the development of "prepackaged" group identities such as nationalities and ethnic groups through our evolutionary history as well as represent how specific identities emerge and shift among contemporary populations. We propose that extending existing computational models of categorization and concept learning (e.g., Love, Medin, & Gureckis, 2004) into the social domain by including dimensions such as identity or motivation may offer a highly tractable account for how we can deploy basic, domain-general inference processes to solve the challenges of reasoning about large and complex groups.

Financial support. We acknowledge the support of the Social Sciences and Humanities Research Council of Canada to R.A.G. (SSHRC CGS-M), S.A. (SSHRC CGS-D), and W.A.C. (SSHRC-506457), and the support of the V. Kann Rasmussen Foundation to D.H.

Conflict of interest. None.

References

Baron, A. S., Dunham, Y., Banaji, M., & Carey, S. (2014). Constraints on the acquisition of social category concepts. *Journal of Cognition and Development, 15*(2), 238–268. https://doi.org/10.1080/15248372.2012.742902.

Butler, L. P., & Markman, E. M. (2014). Preschoolers use pedagogical cues to guide radical reorganization of category knowledge. *Cognition, 130*(1), 116–127. https://doi.org/10.1016/j.cognition.2013.10.002.

Chalik, L., & Rhodes, M. (2018). Learning about social category-based obligations. *Cognitive Development, 48,* 117–124. https://doi.org/10.1016/j.cogdev.2018.06.010.

Chalik, L., & Rhodes, M. (2020). Groups as moral boundaries: A developmental perspective. In *Advances in child development and behavior* (Vol. 58, pp. 63–93). Elsevier. https://doi.org/10.1016/bs.acdb.2020.01.003.

Cunningham, W. A., Van Bavel, J. J., & Johnsen, I. R. (2008). Affective flexibility: Evaluative processing goals shape amygdala activity. *Psychological Science, 19*(2), 152–160. https://doi.org/10.1037/e617962012-341.

Dunbar, R. (2010). *How many friends does one person need?: Dunbar's number and other evolutionary quirks.* Faber & Faber.

Foster-Hanson, E., & Rhodes, M. (2019). Normative social role concepts in early childhood. *Cognitive Science, 43*(8), e12782. https://doi.org/10.1111/cogs.12782.

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences, 114*(23), 6068–6073. https://doi.org/10.1073/pnas.1703857114.

Gelman, S. A., Ware, E. A., & Kleinberg, F. (2010). Effects of generic language on category content and structure. *Cognitive Psychology, 61*(3), 273–301. https://doi.org/10.1016/j.cogpsych.2010.06.001.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*(3), 307–321. https://doi.org/10.1111/j.1467-7687.2007.00585.x.

Lew-Levy, S., Lavi, N., Reckin, R., Cristóbal-Azkarate, J., & Ellis-Davies, K. (2018). How do hunter-gatherer children learn social and gender norms? A meta-ethnographic review. *Cross-Cultural Research, 52*(2), 213–255. https://doi.org/10.1177/1069397117723552.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309.

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology, 66*(1), 37–47.

Roberts, S. O., Gelman, S. A., & Ho, A. K. (2017a). So it is, so it shall be: Group regularities license children's prescriptive judgments. *Cognitive Science, 41*(S3), 576–600. https://doi.org/10.1111/cogs.12443.

Roberts, S. O., Ho, A. K., & Gelman, S. A. (2017b). Group presence, category labels, and generic statements influence children to treat descriptive group regularities as prescriptive. *Journal of Experimental Child Psychology, 158*, 19–31. https://doi.org/10.1016/j.jecp.2016.11.013.

Roccas, S., & Brewer, M. B. (2002). Social identity complexity. *Personality and Social Psychology Review, 6*(2), 88–106. https://doi.org/10.1207/S15327957PSPR0602_01.
Taverna, A. S., Padilla, M. I., Baiocchi, M. C., & Peralta, O. A. (2021). Collaborative pedagogy: 3-year-olds bring pedagogical cues into alignment with analogical reasoning to extract generic knowledge. *European Journal of Psychology of Education, 36*(2), 423– 438. https://doi.org/10.1007/s10212-020-00475-4.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788.
Van Bavel, J. J., & Cunningham, W. A. (2012). A social identity approach to person memory: Group membership, collective identification, and social role shape attention and memory. *Personality and Social Psychology Bulletin, 38*(12), 1566–1578. https://doi.org/10.1177/0146167212455829.
Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2011). Modulation of the fusiform face area following minimal exposure to motivationally relevant faces: Evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience, 23*(11), 3343–3354. https://doi.org/10.1162/jocn_a_00016.

Paranoia reveals the complexity in assigning individuals to groups on the basis of inferred intentions

Anna Greenburgh and Nichola Raihani

Department of Experimental Psychology, University College London, London WC1H 0AP, UK
a.greenburgh@ucl.ac.uk | n.raihani@ucl.ac.uk | doi:10.1017/S0140525X21001229, e109

Abstract

We suggest that variation, error, and bias will be essential to include in a complete computational theory of groups – particularly given that formation of group representations must often rely on inferences of intentions. We draw on the case study of paranoia to illustrate that intentions that do not correspond to group-constitutive roles may often be perceived as such.

The target article offers a computational theory of groups, suggesting that group membership can be inferred based on the assignment of agents to specific roles within triadic conflict. We appreciate the value in the author's clear conceptualisation of group membership as a relational property. Nevertheless, we wish to raise the overlooked issue of variation, error, and bias when assigning others to roles and, thus, inferring group membership. The target article acknowledges that assignments to roles is inherently probabilistic; we outline how this may often stem from biases in attributing intentions to others, and that this might complicate the ways that individuals understand the social structure of the world they inhabit.

In the absence of directly perceived action, formation of group representations often relies on attributing intentions to others. The target article states that what makes intentions relevant to group representations is whether they can predict group constituent roles [“what makes a particular intention or motivation ‘genuinely’ about groups is that it will lead agents to occupy the group-constitutive roles across all four interaction types, both now and in the future” (note 10)].

However, we argue that this view obscures the reality of how group representations are often formed on the basis of inferring intentions. We suggest that whether intentions are group-based according to the target article's definition (that they lead agents to occupy group-constitutive roles) is often ambiguous. Moreover, intentions that are not group-based can be interpreted as such.

A clear example of biases in inferring others' intentions can be seen in paranoia. Paranoia exists on a spectrum of severity in the general population and need not be indicative of any clinical disorder (Bebbington et al., 2013). Our study demonstrates that paranoia is positively associated with a tendency to attribute malevolent intent to others even when true intentions are ambiguous (Raihani & Bell, 2017; Saalfeld, Ramadan, Bell, & Raihani, 2018). In particular, we find that paranoia involves a lowered threshold for detecting harmful intentions from both cohesive and non-cohesive groups (Greenburgh, Bell, & Raihani, 2019). In other words, we find evidence of a bias to perceive malevolent group-based intentions even when signals for group-constitutive roles are weak.

This perception of malevolent intentions directly pertains to biased group representations in paranoia: Paranoia is commonly characterised by the heightened belief that others are coordinating as a *group intending* to harm the individual (Raihani & Bell, 2019). For example, an item in the Revised Green et al.

Paranoid Thoughts Scale that highly discriminates shifts in paranoia in the general population is “I was convinced there was a conspiracy against me,” and endorsement of this item is a strong indicator of heightened paranoia (Freeman et al., 2021). Therefore, from the perspective of triadic interactions described in the target article, paranoia typically involves skewed group representations: Paranoid individuals often detect conspiracies reminiscent of “alliance”-type conflicts (Fig. 2 in the target article) even when the individuals involved in these conspiracies may not be part of a coherent group with coordinated aims in reality.

Another common example of variation in group representation is provided by conspiracy thinking – which is a distinct but correlated construct to paranoia (Imhoff & Lamberty, 2018). Conspiracy thinking refers to the belief that significant public events are caused by secret plots by two or more powerful, and often malevolent, actors (Douglas et al., 2019). Conspiracy thinking is widespread but variable in the general population (Freeman et al., 2020; Freeman & Bentall, 2017), providing another example of how, when group-constitutive information is ambiguous, some individuals can form group representations.

At the extremity of the paranoia continuum, in persecutory delusions, group perception can arise in the absence of any group-constitutive information at all. The target article suggests that the cognitive system attends to intentions as sources of group-based information when intentions predict whether agents will occupy group-constitutive roles in all interaction types, both at present and in the future. However, this is not the case where persecutory delusions are concerned – persecutory delusions are often characterised by the perception of a conspiracy organised to target the individual, even though no group with such intentions necessarily exists in the material world (Cameron, 1959), and therefore these beliefs have no predictive value for future group-constitutive roles. For example, Green et al. (2006) report that 81.7% of a sample of individuals with current persecutory delusions believed their persecutors were organised into a conspiracy against them.

Given these known biases in inferring malevolent intentions, any computational model of groups must be able to allow for variation in how group representations are formed. Answers to the question posed by the target article, “What is a group?,” will vary significantly between humans.

Financial support. Both authors receive funding from The Royal Society.

Conflict of interest. None.

References

Bebbington, P. E., McBride, O., Steel, C., Kuipers, E., Radovanoviĉ, M., Brugha, T., … Freeman, D. (2013). The structure of paranoia in the general population. *The British Journal of Psychiatry, 202*(6), 419–427.
Cameron, N. (1959). The paranoid pseudo-community revisited. *American Journal of Sociology, 65*(1), 52–58.
Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology, 40*, 3–35.
Freeman, D., & Bentall, R. P. (2017). The concomitants of conspiracy concerns. *Social Psychiatry and Psychiatric Epidemiology, 52*(5), 595–604.
Freeman, D., Loe, B. S., Kingdon, D., Startup, H., Molodynski, A., Rosebrock, L., … Bird, J. C. (2021). The revised Green et al., Paranoid Thoughts Scale (R-GPTS): Psychometric properties, severity ranges, and clinical cut-offs. *Psychological Medicine, 51*(2), 244–253.
Freeman, D., Waite, F., Rosebrock, L., Petit, A., Causier, C., East, A., … Lambe, S. (2020). Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological Medicine, 52*(2), 1–13. doi: 10.1017/S0033291720001890.
Greenburgh, A., Bell, V., & Raihani, N. (2019). Paranoia and conspiracy: Group cohesion increases harmful intent attribution in the trust game. *PeerJ, 7,* e7403.
Green, C., Garety, P. A., Freeman, D., Fowler, D., Bebbington, P., Dunn, G., & Kuipers, E. (2006). Content and affect in persecutory delusions. *British Journal of Clinical Psychology, 45,* 561–577. doi: 10.1348/014466506X98768.
Imhoff, R., & Lamberty, P. (2018). How paranoid are conspiracy believers? Toward a more fine-grained understanding of the connect and disconnect between paranoia and belief in conspiracy theories. *European Journal of Social Psychology, 48*(7), 909– 926.
Raihani, N. J., & Bell, V. (2017). Paranoia and the social representation of others: A largescale game theory approach. *Scientific Reports, 7*(1), 1–9.
Raihani, N. J., & Bell, V. (2019). An evolutionary perspective on paranoia. *Nature Human Behaviour, 3*(2), 114–121.
Saalfeld, V., Ramadan, Z., Bell, V., & Raihani, N. J. (2018). Experimentally induced social threat increases paranoid thinking. *Royal Society Open Science, 5*(8), 180569.

Compassion within conflict: Toward a computational theory of social groups informed by maternal brain physiology

S. Shaun Ho[a], Richard N. Rosenthal[a], Helen Fox[a], David Garry[b], Meroona Gopang[a], Mikaela J. Rollins[a], Sarah Soliman[a] and James E. Swain[a,b,c]

[a]Department of Psychiatry and Behavioral Health, Renaissance School of Medicine, Stony Brook University, Stony Brook, NY 11794, USA; [b]Department of Obstetrics, Gynecology & Reproductive Medicine, Renaissance School of Medicine, Stony Brook University, Stony Brook, NY 11794, USA and [c]Department of Psychology, Renaissance School of Medicine, Stony Brook University, Stony
Brook, NY 11794, USA
Shao-Hsuan.ho@stonybrookmedicine.edu | mikaelajrollins@email.arizona.edu | sarah.soliman@stonybrook.edu | meroona.gopang@stonybrook.edu | Helen.Fox@stonybrookmedicine.edu | David.Garry@stonybrookmedicine.edu | Richard.Rosenthal@stonybrookmedicine.edu | James.swain@stonybrookmedicine.edu;
https://www.stonybrookmedicine.edu/profile?pid=2038&name=James%20Swain%20MD | doi:10.1017/S0140525X21001436, e110

Abstract

Benevolent intersubjectivity developed in parent–infant interactions and compassion toward friend and foe alike are nonviolent interventions to group behavior in conflict. Based on a dyadic active inference framework rooted in specific parental brain mechanisms, we suggest that interventions promoting compassion and intersubjectivity can reduce stress, and that compassionate mediation may resolve conflicts.

Computational theories of n-person conflict aim to explain triadic interactions between agents A, B, and C, in which C is drawn into dyadic conflicts between A and B in one of four scenarios: generalization, alliance, defense, and displacement (Pietraszewski, target article). Although Pietraszewski considered A and C forming a group in the displacement scenario, in which A attacks B, and B attacks C in a chain of attacks, we would argue that there is no group in this type, as the relationship between A and C is undetermined, which can be either (1) unrelated, (2) allied (so B attacks C to retaliate against A), or (3) indirectly antagonistic (A attacks B to cause B to attack C). Furthermore, we propose an additional scenario, namely compassionate mediation, in which C prevents A from attacking B in any of multiple ways, for example, C can deter A from attacking B by fortifying B's defense, or C can eliminate A's aggression by providing an alternative solution to A's problem that causes A to attack B. As the compassion here requires equanimity with friends and foes, with a genuine intention to relieve their suffering equally (Ho, Nakamura, & Swain, 2020b), such compassion mediation is fostered by benevolent intersubjectivity, that is, attuning to self and other's needs with benevolence, which can be modeled in a dyadic active inference framework (Ho et al., 2020a), described below. We postulate that if C can first establish dyadic intersubjectivity with A and B separately, that is, C–A and C–B, and if C intends to use dyadic intersubjectivity to address A and B's intentions and/or needs equally, then C can accomplish the compassionate mediation between A and B (A–C–B). Although A and B would never be in the same group in Pietraszewski's conflict model, A, B, and C would form a group altogether as a result of compassionate mediation.

As the building blocks of compassion mediation, we summarize the dyadic active inference framework (Ho et al., 2020a) as follows: (1) Each agent has an active inference engine, in which there is an two-node interface of afferent feelings and efferent actions, along with an internal model node to infer the other agent's internal models; (2) if the internal model fails to reduce its prediction errors the agent will perceive stress; (3) when one agent's feelings are caused by the other's actions and vice versa, they are strongly coupled; (4) normally, such strong coupling enables the agents to reduce prediction errors of each other's internal models and thus reduce stress; (5) undercoupling will cause any agent to ignore the impact of their actions on another agent and thus results in failures to understand the other agent's internal model; (6) stress can cause over-mentalizing of others by preventing an agent from updating their internal model because of unresolvable conflicts; and (7) holding on to ineffective internal models despite failures to minimize prediction errors can result in a vicious cycle of over-mentalizing and undercoupling and exacerbate conflict.

Thus, modeling group conflict may include the active inference engine as cognitive primitives underlying the undercoupling and over-mentalizing problems described above. Here, we discuss a brain model of an active inference engine, based on the neural responses during maternal "mirroring" of the child's feelings and actions, which covaried with parenting stress reduction after an intersubjectivity-promoting parenting intervention (Ho et al., 2020a): When the mother mirrored the child's facial expressions, such valid and genuine mirroring deactivated the default-mode network (which would mediate the internal models of the mother's engine) and, conversely, activated the mirror neuron system (which would mediate the feeling/action interface of the mother's engine) and the salience network (which would be triggered by prediction errors).

The salience network plays a pivotal role in the postconflict adaptation. After a conflict triggers an agent's salience network, his or her compassionate mediation may be enabled if the agent can suspend the previous internal model in the default-mode network and engage in the mirror-neuron system to understand other agents more genuinely; otherwise, defensive reactions, rather than compassionate mediation, when the postconflict salience network activates the default-mode network, resulting in excessive perseverance of the failing internal model and thus over-mentalizing of others (Ho et al., 2020b). The importance of the salience network after conflicts is corroborated as the impairment of the salience network is common to psychopathology, including substance use disorders (Goodkind et al., 2015).

The salience network overlaps with maternal behavior neurocircuit (MBN) that regulates the balance between aggression and care in the maternal brain (Swain & Ho, 2017; Swain, Ho, Fox, Garry, & Brummelte, 2019). Indeed, the MBN mediates sensitive parenting in infant development to becoming compassionate agents themselves (Ainsworth, Blehar, Waters, & Wall, 1978; Elmadih et al., 2016; Kim, Strathearn, & Swain, 2016; Kim et al., 2015b; Mayes, Swain, & Leckman, 2005). The MBN, thus, contains brain systems critical to conflict resolution (Eslinger et al., 2021; Guo, Moses-Kolko, Phillips, Swain, & Hipwell, 2018; Hipwell, Guo, Phillips, Swain, & Moses-Kolko, 2015; Swain, 2011; Swain, Kim, & Ho, 2011; Swain & Lorberbaum, 2008; Swain, Lorberbaum, Kose, & Strathearn, 2007), which can be modeled as adversely affected by psychosocial stressors and psychopathology (Ho & Swain, 2017; Kim, Ho, Evans, Liberzon, & Swain, 2015a; Moses-Kolko, Horner, Phillips, Hipwell, & Swain, 2014; Pawluski, Swain, & Lonstein, 2021; Swain et al., 2017; Swain & Ho, 2019, 2021). Adaptive parent–child dyadic interactions and parent–parent–child or parent–child–child triadic interactions may shape the salience network (equivalent to MBN) in participating agents, such that they are more likely to employ compassionate mediation in the context of conflicts.

We hope to see a computational model that can explain all types of scenarios in which agent C may exert violent or nonviolent interventions in the context of conflicts. Future computational models of conflict may consider a triadic active inference framework to explain agent C's participation in terms of how the agents' active inference engines are coupled with one another.

Acknowledgments. This study has been supported by the Research Foundation for the State University of New York (SUNY) and the National Institutes for Health (NIH): National Center for Advanced Translational Sciences (NCATS) via the Michigan Institute for Clinical Health Research UL1TR000433, and National Institute on Drug Abuse (NIDA) R01 DA047336 and DA047336-02S1

Conflict of interest. All authors of the manuscript participated in writing this manuscript and have no conflicts of interest to declare.

References

Ainsworth, M. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation.* Erlbaum.

Elmadih, A., Wan, M. W., Downey, D., Elliott, R., Swain, J. E., & Abel, K. M. (2016). Natural variation in maternal sensitivity is reflected in maternal brain responses to infant stimuli. *Behavioral Neuroscience, 130*(5), 500–510. doi: 10.1037/bne0000161

Eslinger, P. J., Anders, S., Ballarini, T., Boutros, S., Krach, S., Mayer, A. V., … Zahn, R. (2021). The neuroscience of social feelings: Mechanisms of adaptive social functioning. *Neuroscience & Biobehavioral Reviews, 128,* 592–620. doi: 10.1016/j.neubiorev.2021.05.028

Goodkind, M., Eickhoff, S. B., Oathes, D. J., Jiang, Y., Chang, A., Jones-Hagata, L. B., … Etkin, A. (2015). Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry, 72*(4), 305–315. doi: 10.1001/jamapsychiatry.2014.2206

Guo, C., Moses-Kolko, E., Phillips, M., Swain, J. E., & Hipwell, A. E. (2018). Severity of anxiety moderates the association between neural circuits and maternal behaviors in the postpartum period. *Cognitive Affective & Behavioral Neuroscience, 18*(3), 426– 436. doi: 10.3758/s13415-017-0516-x

Hipwell, A. E., Guo, C., Phillips, M. L., Swain, J. E., & Moses-Kolko, E. L. (2015). Right frontoinsular cortex and subcortical activity to infant cry is associated with maternal mental state talk. *Journal of Neuroscience, 35*(37), 12725–12732. doi: 10.1523/JNEUROSCI.1286-15.2015

Ho, S. S., Muzik, M., Rosenblum, K. L., Morelen, D., Nakamura, Y., & Swain, J. E. (2020a). Potential neural mediators of mom power parenting intervention effects on maternal intersubjectivity and stress resilience. *Frontiers in Psychiatry, 11*, 568824. doi: 10.3389/fpsyt.2020.568824

Ho, S. S., Nakamura, Y., & Swain, J. E. (2020b). Compassion as an intervention to attune to universal suffering of self and others in conflicts: A translational framework. *Frontiers in Psychology, 11,* 603385. doi: 10.3389/fpsyg.2020.603385

Ho, S. S., & Swain, J. E. (2017). Depression alters maternal extended amygdala response and functional connectivity during distress signals in attachment relationship. *Behavioural Brain Research, 325*(Pt B), 290–296. doi: 10.1016/j.bbr.2017.02.045

Kim, P., Ho, S. S., Evans, G. W., Liberzon, I., & Swain, J. E. (2015a). Childhood social inequalities influences neural processes in young adult caregiving. *Developmental Psychobiology, 57*(8), 948–960. doi: 10.1002/dev.21325

Kim, P., Rigo, P., Leckman, J. F., Mayes, L. C., Cole, P. M., Feldman, R., & Swain, J. E. (2015b). A prospective longitudinal study of perceived infant outcomes at 18–24 months: Neural and psychological correlates of parental thoughts and actions assessed during the first month postpartum. *Frontiers in Psychology, 6,* 1772. doi: 10.3389/fpsyg.2015.01772

Kim, P., Strathearn, L., & Swain, J. E. (2016). The maternal brain and its plasticity in humans. *Hormones and Behavior, 77*, 113–123. doi: 10.1016/j.yhbeh.2015.08.001

Mayes, L. C., Swain, J. E., & Leckman, J. F. (2005). Parental attachment systems: Neural circuits, genes, and experiential contributions to parental engagement. *Clinical Neuroscience Research, 4*(5–6), 301–313. doi: 10.1016/j.cnr.2005.03.009.

Moses-Kolko, E. L., Horner, M. S., Phillips, M. L., Hipwell, A. E., & Swain, J. E. (2014). In search of neural endophenotypes of postpartum psychopathology and disrupted maternal caregiving. *Journal of Neuroendocrinology, 26*(10), 665–684. doi: 10.1111/jne.12183

Pawluski, J. L., Swain, J. E., & Lonstein, J. S. (2021). Neurobiology of peripartum mental illness. *Handbook of Clinical Neurology, 182,* 63–82.

Swain, J. E. (2011). The human parental brain: In vivo neuroimaging. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 35*(5), 1242–1254.

Swain, J. E., & Ho, S. S. (2017). Neuroendocrine mechanisms for parental sensitivity: Overview, recent advances and future directions. *Current Opinion in Psychology, 15,* 105–110. doi: 10.1016/j.copsyc.2017.02.027

Swain, J. E., & Ho, S. S. (2019). Early postpartum resting-state functional connectivity for mothers receiving buprenorphine treatment for opioid use disorder: A pilot study. *Journal of Neuroendocrinology, 31*(9), e12770. doi: 10.1111/jne.12770

Swain, J. E., & Ho, S. S. (2021). Opioids and maternal brain-behavior adaptation. *Neuropsychopharmacology, 46*(1), 265–266. doi: 10.1038/s41386-020-00858-7

Swain, J. E., Ho, S. S., Fox, H., Garry, D., & Brummelte, S. (2019). Effects of opioids on the parental brain in health and disease. *Frontiers in Neuroendocrinology, 54*, 100766. doi: 10.1016/j.yfrne.2019.100766

Swain, J. E., Ho, S. S., Rosenblum, K. L., Morelen, D., Dayton, C. J., & Muzik, M. (2017). Parent–child intervention decreases stress and increases maternal brain activity and connectivity during own baby-cry: An exploratory study. *Development and Psychopathology, 29*(2), 535–553. doi: 10.1017/S0954579417000165

Swain, J. E., Kim, P., & Ho, S. S. (2011). Neuroendocrinology of parental response to baby-cry. *Journal of Neuroendocrinology, 23*(11), 1036–1041.

Swain, J. E., & Lorberbaum, J. P. (2008). Imaging the human parental brain. *Neurobiology of the Parental Brain, 6,* 83–100. doi: 10.1016/B978-0-12-374285-8.00006-8.

Swain, J. E., Lorberbaum, J. P., Kose, S., & Strathearn, L. (2007). Brain basis of early parent–infant interactions: Psychology, physiology, and in vivo functional neuroimaging studies. *Journal of Child Psychology and Psychiatry, 48*(3–4), 262–287.

Learning agents that acquire representations of social groups

Joel Z. Leibo, Alexander Sasha Vezhnevets, Maria K. Eckstein, John P. Agapiou and Edgar A. Duéñez-Guzmán

DeepMind, London EC4A 3TW, UK
jzl@deepmind.com | vezhnick@deepmind.com | mariaeckstein@deepmind.com | jagapiou@deepmind.com | duenez@deepmind.com | www.jzleibo.com | doi:10.1017/S0140525X21001357, e111

Abstract

Humans are learning agents that acquire social group representations from experience. Here, we discuss how to construct artificial agents capable of this feat. One approach, based on deep reinforcement learning, allows the necessary representations to self-organize. This minimizes the need for hand-engineering, improving robustness and scalability. It also enables "virtual neuroscience" research on the learned representations.

The target article argues for an approach to coalitional psychology involving a focus on the question of how a robot could be made to "see" groups (Pietraszewski). We agree that considering representations is central to that objective. We would like, however, to propose a different approach, which nicely complements that of the target article as it is aimed at a different level of analysis.

Philosophically, our approach accords with the extension to Marr's levels of analysis proposed in Poggio (2012). Its key idea is that in addition to Marr's three classical levels, we can also describe a system at the level of the principles of learning needed for the system to self-organize into a solution to the problem. Recent successes in artificial intelligence show that it is possible to solve difficult problems without a computational or algorithmic understanding of how the system comes to solve them (e.g., language models which do not need linguistics, such as Brown et al., 2020). On this level of analysis, it is not the representation itself that needs to be understood, but rather how it can be learned.

Another way in which our approach differs from that of the target article is that we think that the data the brain use to train its representation of groups likely includes action as well as perception. This is because agents must learn representations that not only help them perceive the world, but also act appropriately in it. Thus, we base our approach on deep reinforcement learning (Mnih et al., 2015): A framework where agents receive observations of their environment and process them into internal representations, which they then use to select actions. The environment has a state (which is often known only imperfectly by the agent). Agents receive rewards when they take certain actions in certain states, and learning proceeds by systematically tweaking representations and decision rules (encoded in a neural network) in order to further an objective of maximizing expected future rewards.

A common misconception of learning-based approaches to cognitive science is that they entail a blank slate perspective wherein data alone induce all the mind's structure. In practice, this could not be further from the truth. For instance, when neural networks are trained to represent natural images, convolutional architectures learn Gabor-like receptive fields (resembling primary visual cortex), but fully connected architectures do not (e.g., Saxe, Bhand, Mudur, Suresh, & Ng, 2011). All network architectures feature inductive bias of one form or another. Therefore, the question always boils down to: What is the right architectural inductive bias to solve the problem at hand?

To capture coalitional psychology, we propose the latent variable model shown in Figure 1. It aims to decompose social interactions into recognizable and reproducible behavioral primitives – called options (Sutton, Precup, & Singh, 1999). The recognition part of the model could be learned in an unsupervised way, for instance by optimizing network weights so as to maximize mutual information between latent variables and observed behavior of other agents, as in Vezhnevets, Wu, Eckstein, Leblond, and Leibo (2020). Some latent variables may come to encode the options of others (e.g., attack and defend), whereas others could be informative ancillary attributes – such as clothing style, language, and so on. The behavior reproduction part of the model could be learned with a method resembling that of Vezhnevets et al. (2017) – the choice of option specifying a desired change in the (social) environment that the network must learn to produce through actions. Decomposing behavior in this way simultaneously induces a representation of the social world (what is going on), and presents a set of implementable options to take in response to it (what to do about it).

The network layers between the recognition and reproduction representations can be interpreted as decision-making circuitry, and are trained by reinforcement learning (e.g., Mnih et al., 2015). These layers should come to represent and use decision-relevant information such as group membership. After learning, the result is an agent that can select appropriate response options for its current context (e.g., encountering ingroup vs. outgroup individuals).

Note that this model dispenses with many of the explicit information-processing functions mentioned in the target article. It does not need any explicit machinery for stringing together chains of triadic primitives, modifiers of defaults, or generating counterfactuals. If these mechanisms are necessary then they will emerge implicitly, just as the perceptual representation and option-production circuitry emerge (Botvinick et al., 2017). By adopting the architecture of Figure 1, the system designer is specifying only that the agent will aim to decompose its social world into options, but not precisely what those options will be, or how to accomplish the decomposition. Instead, the agent learns all that for itself, using the data generated from its interactions with other individuals in its environment. In engineering terms, this kind of approach is thought to be more robust and scalable than one that relies on explicit engineering of each information processing function (LeCun, Bengio, & Hinton, 2015).

Once we have such a learning agent then we can study it *in silico* with neuroscience-inspired analysis methods (e.g., Zhuang et al., 2021). The key question of where and how the agent comes to represent social group assignments becomes one that we can answer empirically. In particular, one could vary the social environment (e.g., conflict vs. cooperation) and probe how different representations emerge as a result. Methods such as representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008) can be applied to explicitly compare these emergent representations to the explicit representation proposed in the target article.

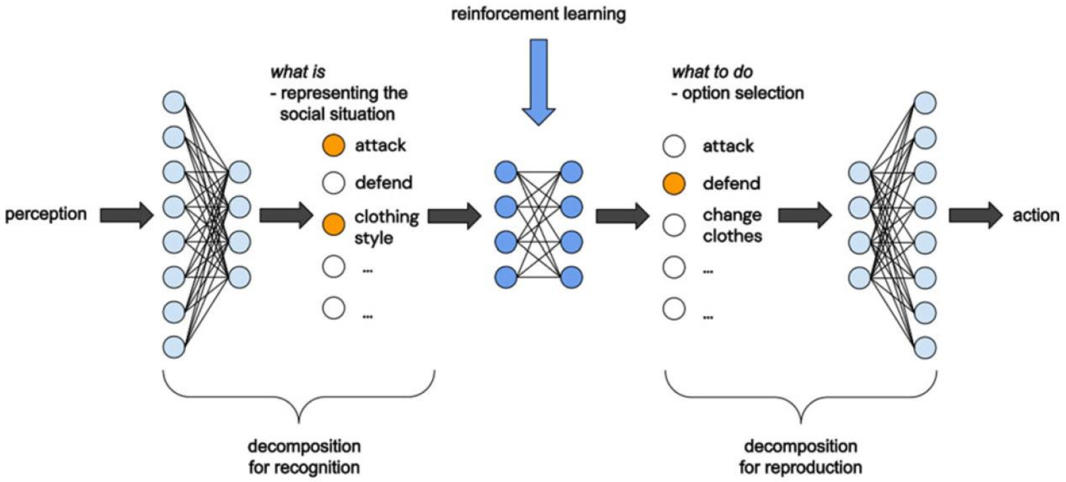


Figure 1 (Leibo et al.). Proposed architecture for a learning agent that can acquire social group representations from experience.

Modern deep reinforcement learning methods enable an alternative approach to the question of how agents represent social groups. Where the target article explicitly enumerates information processing functions, the approach we propose instead involves neural networks that self-organize to solve reinforcement learning problems and softer forms of inductive bias. This level of understanding complements the more explicit level described in the target article.

Financial support. All authors are employees of DeepMind.

Conflict of interest. None.

References

Botvinick, M., Barrett, D. G., Battaglia, P., de Freitas, N., Kumaran, D., Leibo, J. Z., (2017). Building machines that learn and think for themselves [Commentary on Lake et al.] *Behavioral and Brain Sciences, 40,* e255. https://doi.org/10.1017/S0140525X17000048

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … (2020). Language models are few-shot learners. *arXiv preprint arXiv,* 2005.14165.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis – Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2,* 4.

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … (2015) Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

Poggio, T. (2012). The levels of understanding framework, revised. *Perception, 41*(9), 1017–1023.

Saxe, A. M., Bhand, M., Mudur, R., Suresh, B., & Ng, A. Y. (2011). Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. *Advances in Neural Information Processing Systems*, 1971–1979.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence, 112*(1–2), 181–211.

Vezhnevets, A., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., & Kavukcuoglu, K. (2017). Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning,* 3540–3549. PMLR.

Vezhnevets, A., Wu, Y., Eckstein, M., Leblond, R., & Leibo, J. Z. (2020). Options as responses: Grounding behavioural hierarchies in multi-agent reinforcement learning. In *International Conference on Machine Learning,* 9733–9742. PMLR.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences, 118*(3), e2014196118. https://doi.org/10.1073/pnas.2014196118.

Conciliation and meta-contrast are important for understanding how people assign group memberships during conflict situations

Mark Levine and Richard Philpot

Department of Psychology, Lancaster University, Lancaster LA1 4YF, UK
mark.levine@lancaster.ac.uk | r.philpot@lancaster.ac.uk | https://www.lancaster.ac.uk/people-profiles/mark-levine |
https://www.lancaster.ac.uk/psychology/about-us/people/richard-philpot | doi:10.1017/S0140525X2100131X, e112

Abstract

Pietraszewski misrepresents both the nature of behaviour in conflict and the ability of psychology to theorise the relational properties of group designation. At the behavioural level, he focusses exclusively on “attack,” when consolation/care in conflict is equally present and important. At the theoretical level, he ignores existing psychological work on how group perception is shaped by the meta-contrast principle.

In the target article, Pietraszewski argues for a computational approach to deriving how humans assign membership of social groups – and does so using “behaviour in conflict” as the context in which the principles of a computational theory of social groups can be established. Key to the approach is the importance of third-party involvement in conflict (“triads not dyads”) which, in turn, provides the foundation for four “cognitive primitives.” These cognitive primitives are presented as the building blocks of the computational approach. Each cognitive primitive is structured around the possible combinations for any of the three parties to attack each other in turn. Pietraszewski argues that the way in which people interpret these attacking moves then structures how group memberships are assigned.

The major challenge for any computational model is its relationship to a “ground truth.” Thus, Pietrazewski’s approach needs to be measured against what we know about the nature of human behaviour in real-life conflict. There is now a significant literature which examines conflict between humans captured on public CCTV cameras. For example, early work by Levine, Taylor, and Best (2011) used CCTV footage to explore the role of third parties in the escalation and de-escalation of aggression and violence in public space. More recently, work by Liebst, Philpot, and colleagues (Liebst, Philpot, Levine, & Lindegaard, 2021; Philpot, Liebst, Levine, Bernasco, & Lindegaard, 2020a) has explored the behaviours of third parties to public violence, and the likelihood and consequences of such intervention. This work confirms the importance of the triadic approach, but tells an importantly different story about the nature of human behaviour in conflict. Although Pietraszewski focusses exclusively on the propensity for agents to attack each other, systematic behavioural analysis shows that conflict behaviours are a mix of escalation and de-escalation (Ejbye-Ernst, Lindegaard, & Bernasco, in press; Liebst et al., 2019). Moreover, it’s clear from the literature that third parties are much more likely to contribute the latter than the former (Levine et al., 2011; Philpot, 2017). It seems, therefore, that Pietrazewski’s exclusive focus on “attack” as the key communicative act in these triadic relationships ignores the richness of human behaviour in conflict. He is in danger of ignoring the “equally old heritage of countermeasures that protect cooperative arrangements against the undermining effects of competition” (de Waal, 2000, p. 590). Behaviours which are aimed at conflict reduction are equally as “primitive” as those that seek power and dominance. In fact, there are good grounds to argue that behaviours which indicate conciliation and care are likely to be as diagnostic (if not more diagnostic) of group relationships in conflicts (Liebst et al., 2019; Philpot, 2017; Philpot, Liebst, Lindegaard, Verbeek, & Levine, 2020b). People watching others in conflict – and people engaging in conflict themselves – are exposed to a more complex sequence of aggressive and conciliatory acts from the antagonists and third parties to conflict than Pietraszewski allows. In short, the model’s claim to have isolated a defined set of “cognitive primitives” is undermined by this overemphasis on “attack, attack, attack.”

In addition to this misrecognition of the nature of behaviour in conflict, we also argue that Pietraszewski fails to adequately acknowledge where theoretical work in social psychology can contribute to his project. More specifically, we take issue with the claim that traditional social psychology approaches to the group are conceptually or practically blind to the relational property of group membership (sect. 8.2. para. 2). For example, the social identity approach (SIA) (Haslam, 2004; Reicher, 2004) draws extensively on the idea of the “meta-contrast principle” (Turner, Hogg, Oakes, Reicher, & Wetherell, 1987) – which is explicitly relational. The meta-contrast principle states that a collection of individuals tend to be categorised as a group to the degree inter alia that the perceived differences between them are less than the perceived differences between them and other people (outgroups) in the comparative context (Haslam, Reicher, & Levine, 2012; Smith & Hogg, 2008). As part of the work on self-categorisation theory, Turner and colleagues have adapted the classic work of Bruner (1957) on the importance of categorisation for the way an individual makes sense of perceptual stimuli in the world. They show that decisions on when individuals are perceived as groups can be subject to the same kinds of relational categorisation processes. It is true there have been few attempts to build this dynamic group formation idea into computational approaches to modelling group processes (but see Philpot, 2017; Salzarulo, 2004, 2006 for work that provides an entry point for theoretical integration). A computational approach to the perceptual mechanics of group formation would be better served by constructively engaging with rather than ignoring or misrepresenting relevant work in social psychology.

In conclusion, the strength of the approach proposed in this paper is that it seeks to model relationality across triadic rather than dyadic relationships. However, at a behavioural level, the approach needs to recognise the central (and equally “primitive”) role of conciliation and care as an indicator of group belongingness in the context of conflict. This would facilitate a more veridical mapping of what actually happens in conflict. It would also assist Pietraszewski stated aim of making this kind of computational approach to group designation generalisable to contexts other than conflict. The paper would also benefit by engaging constructively with the theoretical work in social psychology on relationality in how group membership is derived. An examination of the meta-contrast principle might be useful in modelling how aggressive and conciliatory acts across triadic sequences can result in the emergence of group properties.

Financial support. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Citizen Forensics project (EP/R013144/1 and EP/R033862/1).

Conflict of interest. Mark Levine and Richard Philpot declare that they have no conflicts of interest to disclose.

References

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64*(2), 123–152. https://doi.org/10.1037/h0043805

de Waal, F. B. (2000). Primates – A natural heritage of conflict resolution. *Science, 289*(5479), 586–590. https://doi.org/10.1126/science.289.5479.586

Ejbye-Ernst, P., Lindegaard, M. R., & Bernasco, W. (in press). A CCTV-based analysis of target selection by guardians intervening in interpersonal conflicts. *European Journal of Criminology,* 1–20. https://doi.org/10.1177/1477370820960338

Haslam, S. A. (2004). *Psychology in organizations.* Sage.

Haslam, S. A., Reicher, S. D., & Levine, M. (2012). When other people are heaven, when other people are hell: How social identity determines the nature and impact of social support. In J. Jetten, C. Haslam, & S. A. Haslam (Eds.), *The social cure: Identity, health, and well being* (pp. 157–174). Psychology Press.

Levine, M., Taylor, P. J., & Best, R. (2011). Third parties, violence, and conflict resolution: The role of group size and collective action in the microregulation of violence. *Psychological Science, 22*(3), 406–412. https://doi.org/10.1177/0956797611398495

Liebst, L. S., Philpot, R., Bernasco, W., Dausel, K. L., Ejbye-Ernst, P., Nicolaisen, M. H., & Lindegaard, M. R. (2019). Social relations and presence of others predict bystander intervention: Evidence from violent incidents captured on CCTV. *Aggressive Behavior, 45*(6), 598–609. https://doi.org/10.1002/ab.21853

Liebst, L. S., Philpot, R., Levine, M., & Lindegaard, M. R. (2021). Cross-national CCTV footage shows low victimization risk for bystander interveners in public conflicts. *Psychology of Violence, 11*(1), 11–18. https://doi.org/10.1037/vio0000299

Philpot, R. (2017). *Beyond the dyad: The role of groups and third-parties in the trajectory of violence.* Open Research Exeter, University of Exeter.

Philpot, R., Liebst, L. S., Levine, M., Bernasco, W., & Lindegaard, M. R. (2020a). Would I be helped? Cross-national CCTV footage shows that intervention is the norm in public conflicts. *American Psychologist, 75*(1), 66–75. https://doi.org/10.1037/amp0000469

Philpot, R., Liebst, L. S., Lindegaard, M. R., Verbeek, P., & Levine, M. (2020b). Reconciliation in human adults: A video-assisted naturalistic observational study of

post conflict conciliatory behaviour in interpersonal aggression. *PsyArXiv*. https://doi.org/10.31234/osf.io/9e4rf
Reicher, S. (2004). The context of social identity: Domination, resistance, and change. *Political Psychology, 25*(6), 921–945. https://doi.org/10.1111/j.1467-9221.2004.00403.x
Salzarulo, L. (2004). Formalizing self-categorization theory to simulate the formation of social groups. In C. Hernández, A. López-Paredes, J. Pajares, & J. M. Galán (Eds.), *Proceedings of the 2nd International Conference of European Social Simulation Association*. University of Valladolid.
Salzarulo, L. (2006). A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies and Social Simulation, 9*(1), 1–13.
Smith, J. R., & Hogg, M. A. (2008). Social identity and attitudes. In W. D. Crano, & R. Prislin (Eds.), *Attitudes and attitude change* (pp. 337–360). Psychology Press.
Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory* (pp. x, 239). Basil Blackwell.

Societies and other kinds of social groups

Mark W. Moffett

Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA
naturalist@erols.com | doi:10.1017/S0140525X21001345, e113

Abstract

People live in distinct groups, notably territory-holding societies, whose boundaries aren't neatly defined by the traits that Pietraszewski describes for his socially aligned groups (or SAGs), as I propose calling them, which occur both within and between our societies. Although studying SAGs could prove enlightening, societies are essential human groups that likely existed long before the complex SAGs of today.

David Hume wrote, "The chief obstacle … to our improvement in the moral or metaphysical sciences is the obscurity of the ideas, and ambiguity of the terms." True, but certain terms can be useful because of their vagueness. Hence, any piece of English writing will likely employ the word "group" in a myriad of ways – even to refer, as Pietraszewski says, to people "waiting around for a bus." Although Pietraszewski's ideas intrigue, my concern is, first, with the author's assertion that such a commonplace term (one psychologists have no alternative to using in all sorts of contexts) can be redefined to narrowly serve an academic purpose. What this means is that the groups described by Pietraszewski need a moniker of their own; I propose calling them Pietraszewski groups or socially aligned groups (SAGs).

Another difficulty for Pietraszewski's usage gaining traction except among a few specialists is that a succinct way of expressing what he has in mind seems hard to come by: He doesn't distill a definition of SAGs in the abstract, where one might be expected, but rather buries the lede by only laying it out in an extended passage that begins 3,000 words into his discussion.

My chief argument with Pietraszewski's article, however, is that the word "group" is equated with cooperation (and then in the context of conflict). He's writing from a prevalent perspective that sees human sociality in terms of coalitions and strategic alliances; for example, a society is defined as a "group organized in a cooperative fashion," in the classic Sociobiology by my mentor Wilson (1975). The evolution of cooperation is an important research field, yet I've argued (Moffett, 2013) that for territory-controlling societies – that is, groups of well-defined memberships that remain stable over generations, which in one form or another have always been a fixture of human life (Moffett, 2020) – too much emphasis can be placed on cooperation.

Thinking of sociality in this way comes easy for evolutionary psychologists given how central cooperation is to human survival. The difficulty with seeing societies as based on cooperation, though, is that only their positive attributes are thus recognized; their equally significant discord is overlooked. Simmel (1908), one of the founders of sociology, recognized cooperation and conflict as inextricable "forms of sociation," each inconceivable without the other. Societies contain shifting tapestries of positive and negative interactions that fluctuate widely, depending on current social stresses and such factors as familiarity among the members and their relative social status. Instead of neatly defining, and separating, the societies themselves, SAGs exist in numerous forms both within and between societies and may change fluidly even while society borders remain stable.

Like it or not, our worst enemies are probably members of that group we call our nation, which we are nevertheless willing to fight for, even sometimes die for, while a hermit who contributes nothing has as much claim to a passport as community-minded citizens. Furthermore, individual citizens can decline to fight on behalf of their country; a nation's political adversaries can retreat from any hint of collaboration; states such as Venezuela can descend into social chaos yet have citizens that remain patriotic; and, at the same time, distinct societies can band together should trade between them be beneficial (Moffett, 2019, Ch. 1). In short, society memberships generally track poorly with Pietraszewski's "group-constitutive roles."

I see social identity as the foundational feature of societies and the primary basis for distinguishing one society from the next (and in modern societies, formed from centuries of conquests and assimilations, one ethnicity or race from the next). As I've written elsewhere (Moffett, 2019, p. 27), "a society is [best] conceived not as an assembly of cooperators, but as a certain kind of group in which everyone has a clear sense of membership brought about by a lasting shared identity. Membership in societies of humans and other species is a yes/no matter, with ambiguity [e.g., the status of recent immigrants] rare. The prospects for alliances, whether from friendship, family ties, or social obligations, may rank among the paramount adaptive gains of having societies in many species, yet aren't necessary to the equation." Indeed, the advantages of collaboration may be purely accidental ("proto-cooperation": Allee, 1931; Herbert-Read et al., 2016). Societies are obligatory across the portion of the evolutionary tree (clade) humans share with chimpanzees and bonobos, which live in "communities" that primatologists recognize as well-defined over the long haul (despite females emigrating upon reaching adulthood: Wrangham, 1996). These essential groups would have been universal among our ancestors long before intricate SAGs.

The evidence from psychology suggests that "the foundation of the human ability to form useful social categories is in place in infancy" (Liberman, Woodward, & Kinzler, 2017). The purely perceptual recognition of such categories as native-language speakers (Kinzler, Dupoux, & Spelke, 2007) underlies the later-developing (e.g., Pauker, Xu, Williams, & Biddle, 2016) conceptual understanding of those groups as socially relevant

(Charlesworth & Banaji, in press). One imagines that these categories, largely arising through exposure to identifying "markers," serve as frequent reference points to which humans apply their observations of cooperative and antagonistic interactions (Smaldino, 2019) relevant to the emergence of SAGs. Indeed, one hypothesis is that, even as children (Meyer, Roberts, Jayaratne, & Gelman, 2020), we rapidly and automatically (e.g., Ito & Senholzi, 2013) categorize what we see as essentialized human groups, in much the way we distinguish other species (Gil-White, 2001). In instances where group borders are fuzzy, our minds draw from a variety of cues to tidy them up and thereby reduce the ambiguities of what would otherwise be a confusing world (MacLin & MacLin, 2011; Timeo, Farroni, & Maass, 2017).

Pietraszewski asks "what the human mind is representing when it represents a social group," but for certain groups this representation needn't be built on anything more complex, at its core, than how we distinguish tiger from panda, with our fear of the *other* developing, with time and experience, toward the former; whether our minds represent any collection of things as a group – humans included – isn't necessarily determined by calculations around whether, and how, they might cooperate.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Allee, W. C. (1931). *Animal aggregations.* University of Chicago Press.
Charlesworth, T. E. S., & Banaji, M. R. (in press). The development of social group cognition. In D. Carlston, K. L. Johnson, & K. Hugenberg (Eds.), *Oxford Handbook of social cognition* (2nd ed.). Oxford University Press.
Gil-White, F. J. (2001). Are ethnic groups biological "species" to the human brain? *Current Anthropology, 42,* 515–536.
Herbert-Read, J. E., Romanczuk, P., Krause, S., Strömbom, D., Couillaud, P., Domenici, P. ... Krause, J. (2016). Proto-cooperation: Group hunting sailfish improve hunting success by alternating attacks on grouping prey. *Proceedings of the Royal Society B, 283,* 20161671.
Ito, T. A., & Senholzi, K. B. (2013). Us versus them: Understanding the process of race perception with event-related brain potentials. *Visual Cognition, 21,* 1096–1120.
Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences, 104,* 12577–12580.
Liberman, Z., Woodward, A.L., & Kinzler, K. D. (2017). The origins of social categorization. *Trends in Cognitive Sciences, 21*, 556–568.
MacLin, O. H., & MacLin, M. K. (2011). The role of racial markers in race perception and racial categorization. In R. Adams, N. Ambady, K. Nakayama, & S. Shimojo (Eds.), *The science of social vision* (pp. 321–346). Oxford University Press.
Meyer, M., Roberts, S. O., Jayaratne, T. E., & Gelman, S. A. (2020). Children's beliefs about causes of human characteristics: Genes, environment, or choice? *Journal of Experimental Psychology: General.* http://dx.doi.org/10.1037/xge0000751.
Moffett, M. W. (2013). Human identity and the evolution of societies. *Human Nature, 24*, 219–267.
Moffett, M. W. (2019). *The human swarm: How our societies arise, thrive, and fall.* Basic Books.
Moffett, M. W. (2020). Societies, identity, and belonging. *Proceedings of the American Philosophical Society, 164,* 1–9.
Pauker, K., Xu, Y., Williams, A., & Biddle, A. M. (2016). Race essentialism and social contextual differences in children's racial stereotyping. *Child Development, 87,* 1409–1422.
Simmel, G. (1908). *Soziologie. Untersuchungen über die formen der vergesellschaftung.* Duncker & Humblot.
Smaldino, P. E. (2019). Social identity and cooperation in cultural evolution. *Behavioural Processes, 161,* 108–116.
Timeo, S., Farroni, T., & Maass, A. (2017). Race and color: Two sides of one story? Development of biases in categorical perception. *Child Development,* 88, 83–102.
Wilson, E. O. (1975). *Sociobiology: The new synthesis.* Harvard University Press.
Wrangham, R., & Peterson, D. (1996). *Demonic Males: Apes and the origins of human violence.* Houghton Mifflin.

Can group representations based on relational cues warrant the rich inferences typically drawn from group membership?

Katalin Oláh[a,b] and Ildikó Király[a,b,c]

[a]Department of Cognitive Psychology, Institute of Psychology, Eötvös Loránd University, H-1046 Budapest, Hungary; [b]MTA-ELTE Social Minds Research Group, Eötvös Loránd University, H-1046 Budapest, Hungary and [c]Cognitive Development Centre, Central European University, H-1051 Budapest, Hungary
olah.katalin@ppk.elte.hu | kiraly.ildiko@ppk.elte.hu | doi:10.1017/S0140525X21001308, e114

Abstract

Pietraszewski's model – though promising in many respects – needs to be extended so that it can explain the multitude of rich inferences that people draw from group membership. In this commentary, we highlight some facets of group thinking, especially from the field of developmental psychology, that cannot be unambiguously accounted for by a model that is built solely on relational cues.

Pietraszewski's target article provides a fascinating new approach to describing what the mind represents as a "social group" and how such representations are formed. Essentially, the present model proposes that "group membership is a relational property (…), who 'belongs' to what group is borne out of a calculation of the relative relationships among the agents involved" (sect. 8.2, para. 2). While welcoming this new approach, in this commentary, we would like to highlight a few phenomena observed about group psychology (especially, in the field of developmental psychology) for which it is not quite clear how the current model would account. We suggest that either the theory should be extended so that it provides a framework for interpreting these phenomena as well or the limitation of scope should be made more explicit.

Specifically, representations generally appear to be more elaborate than merely involving information about specific roles in certain social interactions: Ample evidence in psychology suggests that these representations of social groups are conceptually rich not only in the minds of adults, but in those of very young children as well (Liberman, Woodward, & Kinzler, 2017). In fact, perceived group membership allows even young children to draw inferences not only about how people will relate to each other (Rhodes & Chalik, 2013), but also about, for example, what knowledge (e.g., Liberman, Gerdin, Kinzler, & Shaw, 2020; Soley, 2019) or preferences (Shutts, Kinzler, McKee, & Spelke, 2009) they possess. Moreover, even young children are selective in what kind of inferences they draw from different group memberships. For example, they expect friends to share knowledge of personal affairs, while they expect members of a cultural group to share knowledge of cultural norms (e.g., Liberman et al., 2020). Importantly, these inferences seem to arise as early as 12 months of age (Shutts et al., 2009) and based on cues that are not presented to children in the context of social relations (e.g., language). These examples illustrate that from very early on, humans apply information arising from perceived group membership for predicting the behavior of others not specifically in intergroup conflict situations, but rather in preparation to engage in – mostly collaborative – interactions with fellow ingroups. Although it is acknowledged in the paper that the relational model of groups should be extended to contexts other than conflict, evidence in developmental psychology suggests that a sensitivity to behavioral cues that are not necessarily manifested in interpersonal contexts at all (such as similarity in linguistic behavior) may precede the emergence of a sensitivity to relational cues. Although Pietraszewski's model would allow linguistic cues to be considered "ancillary" attributes, it is not evident how 12-month-old infants would come to encode them as such.

Thus, although the model presented in Pietraszewski's target article is promising in describing how the human mind represents social groups and how these representations may be used to predict behavior in specific interpersonal situations, it is not as clear how it can account for inferences that are not pertinent to how people will interact with each other. It is especially challenging for this model to give an account for such group membership-based inferences where the inferences are drawn from the assumption that the individuals are members of the *same group*, and not from a perceived relational contrast in interactions (e.g., generalizing food preferences within, but not between groups or expecting a person to understand a specific language based on knowledge of their group membership).

Pietraszewski distinguishes the cognitive processes taking place during these computations from simple "categorization" as he claims that the latter only speaks to the containment issues (sect. 8.1). However, category representations are, in fact, used to store a large body of knowledge of the given class that can later be used to make predictions of different properties of the tokens. Although this is generally true for any ontological field, one robust phenomenon that seems to be specific to social category representations is that the inferences and generalizations that are drawn from them tend to run wild, resulting in robust stereotypes. It is yet unclear to us, how such group representations allow for this phenomenon to occur. Although the paper gives a very elaborate description of how these group representations are formed, what cues (direct or ancillary) may be taken as input for the computational process, it is less clear how they would feed into further inferences. Even if the model's main goal is to explain the formation of such representations, we believe that these questions should be touched upon, because – especially taking an evolutionary approach – function and computational process cannot be perfectly separated.

We do not claim that these questions would invalidate the presented model, rather, we would like to highlight the need to consider how the abovementioned phenomena relate to the computation model described here. It is possible that "group" representations only refer to those collectives where members are likely to interact with each other in some way and thus, the interaction pattern can predict membership and vice versa. In this case, it would be necessary to consider how these group representations are different from other social "category" representations where intergroup conflict (or even other types of social interactions, such as reciprocity) are not at the core of the category (e.g., "women"). In the presented model, for example, features typical of the collective of "women" would be considered ancillary attributes, whereas we would suggest that conflicts arising between the sexes are on most occasions consequences and not essential features of belonging to this specific social group (or "category"). Possibly, an elaborate differentiation between "social group" and "social category" representations would help to disambiguate the issue at hand; however, as of yet, this is unfortunately missing from the literature.

Financial support. This work was supported by the Momentum Program of the Hungarian Academy of Sciences (PI: Ildikó Király; grant number: LP-2017-17/2017).

Conflict of interest. None.

References

Liberman, Z., Gerdin, E., Kinzler, K. D., & Shaw, A. (2020). (Un)common knowledge: Children use social relationships to determine who knows what. *Developmental Science, 23*(6), e12962.
Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). The origins of social categorization. *Trends in Cognitive Sciences, 21*(7), 556–568.
Rhodes, M., & Chalik, L. (2013). Social categories as markers of intrinsic interpersonal obligations. *Psychological Science, 24*(6), 999–1006.
Shutts, K., Kinzler, K. D., McKee, C. B., & Spelke, E. S. (2009). Social information guides infants' selection of foods. *Journal of Cognition and Development, 10*(1–2), 1–17.
Soley, G. (2019). What do group members share? The privileged status of cultural knowledge for children. *Cognitive Science, 43*(10), e12786.

Shared intentionality and the representation of groups; or, how to build a socially adept robot

Ben Phillips

Department of Philosophy, School of History, Philosophy and Religious Studies, Arizona State University, Tempe, AZ 85281, USA
bsphilli@asu.edu | https://www.bensphillips.com/ | doi:10.1017/S0140525X21001242, e115

Abstract

Pietraszewski provides a compelling case that representations of certain interaction-types are the "cognitive primitives" that allow all tokens of *group-in-conflict* to be represented within the mind. Here, I argue that the folk concept GROUP encodes shared intentions and goals as more central than these interaction-types, and that providing a computational theory of social groups will be more difficult than Pietraszewski envisages.

In defending his stimulating proposal, Pietraszewski does not focus on the role that theory of mind (or "mindreading") plays in guiding applications of the folk concept GROUP. Importantly, though, there are good reasons for thinking that attributions of shared intentionality have a central role to play.

There is an ongoing debate as to what exactly shared intentionality involves. At a minimum, it requires more than the just the possession of common goals and intentions. For example, people simultaneously jogging along a trail may have the same goal of getting fit, but they are not engaging in shared intentionality unless they each harbor an intention of the form, "I intend that *we* jog together" (see Bratman, 1999, Ch. 8; Tomasello, Carpenter, Call, Behne, & Moll, 2005).

There is direct evidence that shared intentionality drives costly decisions to help ingroup members over outgroup members (McClung, Placì, Bangerter, Clément, & Bshary, 2017). There is also evidence that applications of the folk concept GROUP are guided by attributions of shared intentionality. A number of studies have found that entitativity perception – the tendency to regard some aggregates of people as more "groupish" than others – is mediated by judgments concerning shared intentionality (for some recent discussions, see Phillips, 2021a, b). For example, people's impression that an aggregate of individuals constitutes a genuine group is enhanced when they observe these individuals moving in synchrony (Ip, Chiu, & Wan, 2006; Lakens & Stel, 2011; Wilson & Gos, 2019). Importantly, these studies suggest that people only tend to regard synchronous movement as a cue for groupishness when they see it as resulting from shared intentionality.

Research into entitativity perception, therefore, suggests that the folk encode shared intentions and goals as central to the concept GROUP; whereas, they encode certain visible cues, such as synchronous movement, as relatively peripheral (or "ancillary" to use Pietraszewski's term). There are various models of conceptual centrality. The core insight is that a feature, F, is more central to a given concept than feature, G, if G is represented as depending on F more than F depends on G (see Sloman, Love, & Ahn, 1998). Thus, the mediational effects outlined above suggest that the folk encode shared intentions and goals as more central to GROUP than synchronous movements, because they represent the latter as causally depending on the former, but not vice versa.

Pietraszewski notes that detecting group-based intentions helps us to predict whether an agent will participate in one of the triadic interactions with certain others (note 10). Arguably, though, just as the folk encode shared intentions as more central to GROUP than coordinated movements, they also encode shared intentions as more central than the triadic interaction types identified by Pietraszewski. In contexts of conflict and cooperation, agents participate in these sorts of interactions precisely because they share certain intentions and goals with fellow group members. For example, when Germany invaded Belgium in 1914, Britain's declaration of war was predictable, in part, because Britain and Belgium shared a suite of goals and intentions (all enshrined in a treaty). By the same token, consider a case in which we are *not* willing to categorize an aggregate as a genuine group. In Pietraszewski's example, some people are waiting for a bus when a motorist throws a stone at one of them. Suppose each person at the bus stop simultaneously hurls a stone back at the driver. The studies outlined above suggest that the folk will not categorize this aggregate of people as a genuine group unless they see them as sharing an intention of the form, "I intend that *we* attack the driver together."

If what I am suggesting is right, this puts pressure on Pietraszewski's claim that representations of triadic interaction types are the "cognitive primitives" that constitute the folk concept of a group-in-conflict. A number of studies have found that when one feature is encoded as more central to a given concept than another feature, the former is a stronger determinant of categorization decisions (e.g., see Ahn, Kim, Lassaline, & Dennis, 2000). Thus, according to the alternative hypothesis on offer, applications of GROUP – in contexts of both cooperation and conflict are guided by representations of certain triadic interactions, but only insofar as these interactions serve as cues for shared intentionality. This suggests that the task of constructing a computational theory of group cognition is more difficult than Pietraszewski envisages, for it will require no less than a computational theory of mindreading. To put it another way, suppose we were to build Pietraszewski's robot, which can navigate the social world by deploying representations of groups. If our robot cannot attribute shared intentionality to others, it will be left in the dust by its (socially adept) human counterparts.

To conclude, it is worth noting that a full-blown capacity for shared intentionality takes time to develop in humans (Tomasello et al., 2005) – presumably, the capacity to *attribute* shared intentionality takes considerably longer. Nonetheless, there is evidence that infants can track some of the interaction-types identified by Pietraszewski (e.g., see Ting, He, & Baillargeon, 2019). Similarly, the capacity for shared intentionality appears to be largely absent in nonhuman primates (Tomasello et al., 2005). Regardless, some nonhuman primates, such as baboons, are able to track groups as they fluctuate across episodes of conflict (e.g., see Cheney & Seyfarth, 2007). It is possible that young children, as well as some nonhuman primates, track groups-in-conflict by detecting triadic interactions of the sort identified by Pietraszewski. If so, Pietraszewski's account may describe an early developing, phylogenetically ancient, system for detecting groups-in-conflict. This system may output a relatively "thin" concept of groups the possession of which does not require an understanding of shared intentionality. Instead, possessing the thin concept might only require an agent to detect triadic interactions by using low-level perceptual cues (e.g., visible instances of hitting, chasing, etc.). In contrast, the "thick" concept of a group that adults deploy in central cognition appears to encode shared intentions as central, and triadic interactions as more peripheral.

Conflict of interest. I have no known conflict of interest to disclose.

References

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41*(4), 361–416.

Bratman, M. (1999). *Faces of intention: Selected essays on intention and agency*, Cambridge University Press.

Cheney, D. L., & Seyfarth, R. M. (2007). *Baboon metaphysics: The evolution of a social mind*. The University of Chicago Press.

Ip, G. W. M., Chiu, C. Y., & Wan, C. (2006). Birds of a feather and birds flocking together: Physical versus behavioral cues may lead to trait- versus goal-based group perception. *Journal of Personality and Social Psychology, 90,* 368–381.

Lakens, D., & Stel, M. (2011). If they move in sync, they must feel in sync: Movement synchrony leads to attributions of rapport and entitativity. *Social Cognition, 29,* 1–14.

McClung, J., Placì, S., Bangerter, A., Clément, F., & Bshary, R. (2017). The language of cooperation: Shared intentionality drives variation in helping as a function of group membership. *Proceedings of the Royal Society B: Biological Sciences, 284.*

Phillips, B. (2021a). The roots of racial categorization. *Review of Philosophy and Psychology.* https://doi.org/10.1007/s13164-021-00525-w.

Phillips, B. (2021b). Entitativity and implicit measures of social cognition. *Mind & Language.* https://doi.org/10.1111/mila.12350.

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22*(2), 189–228.

Ting, F., He, Z., & Baillargeon, R. (2019). Toddlers and infants expect individuals to refrain from helping an ingroup victim's aggressor. *PNAS, 116*(13), 6025–6034.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675–691. doi: 10.1017/S0140525X05000129.

Wilson, S., & Gos, C. (2019). Perceiving social cohesion: Movement synchrony and task demands both matter. *Perception, 48*(4), 316–329.

Developmental antecedents of representing “group” behavior: A commentary on Pietraszewski’s theory of groups

Anthea Pun and Andrew Baron

Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC V6T 1Z4, Canada
antheacp@psych.ubc.ca | abaron@psych.ubc.ca; https://childdevelopment.psych.ubc.ca | doi:10.1017/S0140525X2100145X, e116

Abstract

Central to Pietraszewski’s theory is a set of group-constitutive roles within four triadic primitives. Although some data from the developmental and biological sciences support Pietraszewski’s theory, other data raise questions about whether similar behavioral expectations hold across various ecological conditions and interactions. We discuss the potential for a broader set of conceptual primitives that support reasoning about groups.

Central to Pietraszewski’s computational theory of groups is a set of group-constitutive roles within four triadic primitives. To be considered part of a group, he argues that individuals are obligated to occupy such roles during conflict. Such building blocks for representing groups, if part of humans’ evolved psychology, could be present ontogenetically and perhaps phylogenetically. Although some data from the developmental and biological sciences support Pietraszewski’s theory, other data raises questions about whether “group membership applies to all of the triadic primitives” (sect. 7, sect. 2) and whether similar behavioral expectations hold across various ecological conditions and interactions. In reviewing this work, we open the door to discuss a broader set of conceptual primitives that support reasoning about groups.

Several studies have examined infants’ capacity to reason about social groups, revealing that they are capable of making inferences about multi-agent conflict, and the types of behaviors that should be directed toward ingroup versus outgroup members (Pun, Birch, & Baron, 2016, 2021; Rhodes, Hetherington, Brink, & Wellman, 2015). For example, when 16-month-old infants witnessed a conflict between two agents from opposing groups, they were more surprised when these agents’ social partners cooperated (instead of conflicted) with one another (Rhodes et al., 2015). This result is consistent with Pietraszewski’s *triadic primitive generalization, in which a conflict between two agents from opposing groups can be extended to another, uninvolved member of a group.*

Research with infants as young as 9 months of age supports the early emergence of the primitives *defense* and *alliance* (Pun, Birch, & Baron, 2021). Specifically, after watching two agents from opposing groups come into conflict, infants expected an ingroup member (that was not part of the initial conflict) to harm an outgroup member (by pushing them off the platform). Similar findings have been observed with nonhuman primates, children, and adults (e.g., De Dreu et al., 2016; Rhodes & Brickman, 2011; Rusch, 2013). This suggests that when defending the interests of the group, individuals may be obligated to help ingroup members (i.e., be allies), even if it requires harming outgroup members

Relatedly, by 6 months of age infants understand that there is “strength in numbers.” After observing two agents compete with one another, infants expected the agent with more group members to prevail over an agent with fewer group members (Pun et al., 2016). To be able to make this inference, infants likely inferred that group members acted as *allies*, banding together to support one another against an opposing group. Consistent with this argument, recall that infants were more surprised when a group member did not provide aid to an ingroup member (Pun et al., 2021).

Interestingly, even though infants did not witness “the initially uninvolved onlooker…behave” (sect. 5, para. 3) in Pun et al. (2016), they were still able to predict the outcome of a competition based on group size. Indeed, similar to many social species, from insects, to lions and nonhuman primates (Batchelor & Briffa, 2011; McComb, Packer, & Pusey, 1994; Wilson, Hauser, & Wrangham, 2001), the capacity to predict the outcome of a competition based on the presence of group members is critical for survival and may reflect part of an evolved psychology that supports reasoning about social groups. Therefore, even if cues to group membership (e.g., spatial temporal cues and moving in synchrony) are considered *ancillary* according to Pietraszewski, they may be sufficient to activate group-constitutive roles of novel groups, even prior to extensive socialization.

Other studies with infants and nonhuman species suggest that variation in ecological conditions (e.g., competition, cooperation, resource acquisition, and predation) may activate different computations for ingroup and outgroup members, that ultimately influences behavior (e.g., Avilés, 2002; Bian, Sloane, & Baillargeon, 2018; Bonner, 1982; Chapman & Teichroeb, 2012; Krause, Ruxton, Ruxton, & Ruxton, 2002; Lindstedt et al., 2018). This raises the possibility that any computational theory of groups needs to account for the influence of these conditions. For example, 18-month old infants expect ingroup and outgroup members to be treated equally when resources are abundant. In contrast, ingroup members are expected to be prioritized when resources are limited (Bian et al., 2018), suggesting that competition promotes ingroup favoritism. Relatedly, in the absence of conflict, 17–19-month-olds expect ingroup members to provide instrumental help to another ingroup member (but not an outgroup member) (Jin & Baillargeon, 2017).

Furthermore, ingroup loyalty is expected to be maintained such that behaviors that harm the group are not permissible (Rhodes et al., 2015, Rullo, Presaghi, & Livi, 2015; Ting, He, & Baillargeon, 2019). For example, when ingroup loyalty is violated (e.g., an ingroup member harms another ingroup member), infants in their first year of life expect ingroup members to refrain from helping the aggressor (as a form of punishment) (Ting et al., 2019). However, it is not clear whether committing this transgression leads infants to infer that this disloyal individual should be ostracized from the group. Future research may want to consider the possibility that additional primitives independently shape expectations about ingroups and outgroups.

Finally, research from the biological sciences has proposed that variation in ecological conditions effect the formation and

Originally published in: Behavioral and Brain Sciences, 45, 2022, p. 41

maintenance of groups (Avilés, Fletcher, & Cutter, 2004; Bonner, 1982; Chapman & Teichroeb, 2012; Frank, 2003). For example, in the species *Dictyostelium discoideum* (cellar slime mold), facing resource scarcity and starvation leads to the formation of groups. Ultimately, some of the aggregated cells sacrifice their own reproductive fitness for the survival of the group (Bonner, 1982). This demonstrates that the formation of the group and the roles individuals occupy can vary as a function of the environment (e.g., Avilés, 2002; Krause et al., 2002). Given that these behaviors occur even at the cellular level, it may provide insight into the ways in which group-constitutive roles could be represented psychologically.

Together, research from the developmental and biological sciences provide some initial support for Pietraszewski's theory of groups. This study suggests that individuals may only be obligated to occupy the roles within the triadic primitives during a conflict, potentially because conflict may constrain the behaviors that are permissible for ingroup members to engage in. However, such expectations may not emerge simultaneously in development, even if all of them are part of an evolved psychology.

Furthermore, variation in ecological conditions can influence how individuals within a group should behave toward ingroups and outgroups. For example, when no conflict is present, expectations of behavior may be less rigid; ingroup members may be permitted to behave more benevolently toward outgroup members, and outgroup harm may be less acceptable. Finally, because ingroup support is expected to be maintained, exploring how violating group-constitutive roles affects perceptions of group membership may help inform a more robust computational theory of groups.

Financial support. This research was funded by a Social Sciences and Humanities Research Council Insight Development Grant (Grant No. 435-2013-0286) (to A.S.B.).

Conflict of interest. None.

References

Avilés, L. (2002). Solving the freeloaders paradox: Genetic associations and frequency-dependent selection in the evolution of cooperation among nonrelatives. *Proceedings of the National Academy of Sciences, 99*(22), 14268–14273.

Avilés, L., Fletcher, J. A., & Cutter, A. D. (2004). The kin composition of social groups: Trading group size for degree of altruism. *The American Naturalist, 164*(2), 132–144.

Batchelor, T. P., & Briffa, M. (2011). Fight tactics in wood ants: Individuals in smaller groups fight harder but die faster. *Proceedings of the Royal Society B: Biological Sciences, 278*(1722), 3243–3250.

Bian, L., Sloane, S., & Baillargeon, R. (2018). Infants expect ingroup support to override fairness when resources are limited. *Proceedings of the National Academy of Sciences, 115*(11), 2705–271.

Bonner, J. T. (1982). Evolutionary strategies and developmental constraints in the cellular slime molds. *The American Naturalist, 119*(4), 530–552.

Chapman, C. A., & Teichroeb, J. A. (2012). What influences the size of groups in which primates choose to live. *Nature Education Knowledge, 3*(10), 9.

De Dreu, C. K., Gross, J., Méder, Z., Giffin, M., Prochazkova, E., Krikeb, J., & Columbus, S. (2016). In-group defense, out-group aggression, and coordination failures in intergroup conflict. Proceedings of the National Academy of Sciences, 113(38), 10524–10529. Frank, S. A. (2003). Repression of competition and the evolution of cooperation. *Evolution, 57*(4), 693–705.

Jin, K. S., & Baillargeon, R. (2017). Infants possess an abstract expectation of ingroup support. *Proceedings of the National Academy of Sciences, 114*(31), 8199–8204.

Krause, J., Ruxton, G. D., Ruxton, G., & Ruxton, I. G. (2002). *Living in groups.* Oxford University Press.

Lindstedt, C., Miettinen, A., Freitak, D., Ketola, T., López-Sepulcre, A., Mäntylä, E., & Pakkanen, H. (2018). Ecological conditions alter cooperative behaviour and its costs in a chemically defended sawfly. *Proceedings of the Royal Society B, 285*(1884), 20180466.

McComb, K., Packer, C., & Pusey, A. (1994). Roaring and numerical assessment in contests between groups of female lions, Panthera leo. *Animal Behaviour, 47*(2), 379–387.

Pun, A., Birch, S. A., & Baron, A. S. (2016). Infants use relative numerical group size to infer social dominance. *Proceedings of the National Academy of Sciences, 113*(9), 2376– 2381.

Pun, A., Birch, S. A., & Baron, A. S. (2021). The power of allies: Infants' expectations of social obligations during intergroup conflict. *Cognition, 211*, 104630.

Rhodes, M., & Brickman, D. (2011). The influence of competition on children's social categories. *Journal of Cognition and Development, 12*(2), 194–221.

Rhodes, M., Hetherington, C., Brink, K., & Wellman, H. M. (2015). Infants' use of social partnerships to predict behavior. *Developmental Science, 18*(6), 909–916.

Rullo, M., Presaghi, F., & Livi, S. (2015). Reactions to ingroup and outgroup deviants: An experimental group paradigm for black sheep effect. *PLoS ONE, 10*(5), e0125605.

Rusch, H. (2013). Asymmetries in altruistic behavior during violent intergroup conflict. *Evolutionary Psychology, 11*(5), 973–993.

Ting, F., He, Z., & Baillargeon, R. (2019). Toddlers and infants expect individuals to refrain from helping an ingroup victim's aggressor. *Proceedings of the National Academy of Sciences, 116*(13), 6025–6034.

Wilson, M. L., Hauser, M. D., & Wrangham, R. W. (2001). Does participation in intergroup conflict depend on numerical assessment, range location, or rank for wild chimpanzees?. *Animal Behaviour, 61*(6), 1203–1216.

Triadic conflict "primitives" can be reduced to welfare trade-off ratios

Wenhao Qi, Edward Vul, Adena Schachner and Lindsey J. Powell

Department of Psychology, University of California, San Diego, La Jolla, CA 92093, USA
wqi@ucsd.edu; evul@ucsd.edu; adschachner@ucsd.edu; ljpowell@ucsd.edu; https://jameswhqi.github.io/; https://www.evullab.org/; https://madlab.ucsd.edu/; https://socallab.ucsd.edu/ | doi:10.1017/S0140525X21001382, e117

Abstract

Pietraszewski proposes four triadic "primitives" for representing social groups. We argue that, despite surface differences, these triads can all be reduced to similar underlying welfare trade-off ratios, which are a better candidate for social group primitives. Welfare trade-off ratios also have limitations, however, and we suggest there are multiple computational strategies by which people recognize and reason about social groups.

Pietraszewski convincingly argues for the value of identifying computationally specific principles by which humans recognize and reason about social groups. He then proposes four types of triadic conflict scenarios as computational primitives for representing social groups. Here, we argue that these are not primitives, but are a consequence of, and thus a cue to, a simpler underlying representation: dyadic welfare trade-off ratios.

A welfare trade-off ratio describes how one individual values another's welfare, and thus predicts when they will pay a cost to promote, or detract from, the other's rewards (Delton & Robertson, 2016; Tooby & Cosmides, 2008). In a dyad between person A and person B, if we define A's utility function as $u_{\mathrm{A}} = v_{\mathrm{A}} + \lambda_{\mathrm{AB}} \cdot v_{\mathrm{B}}$, where v_{A} is A's payoff and v_{B} is B's payoff, then λ_{AB} is

A's welfare trade-off ratio toward B. The greater λ_{AB}, the more A values B's payoff, and thus the more likely A is to act to B's benefit, even when those actions are costly or risky; when λ_{AB} is negative, A will be willing to pay a cost to harm B. (For simplicity, in this commentary we assume $\lambda_{\mathrm{AB}} = \lambda_{\mathrm{BA}}$.)

A attacking B provides direct evidence that λ_{AB} is negative. What happens next provides evidence about both the remaining pairwise welfare trade-off ratios in a triad. First, if C attacks B or B attacks C, this indicates λ_{BC} is also negative. Under these circumstances, λ_{AC} will typically be positive. The pressure for this to be true is described by structural balance theory, which holds that equilibrium is achieved when social networks are structured such that the valence of each pairwise connection is consistent with the others. If A and C both benefit from harm to B, then their connection ought to have positive valence (Cartwright & Harary, 1956). (Consider what would happen if all three relations were negative: When C attacks B this would be consistent with λ_{AB} but inconsistent with λ_{AC}, as the attack would indirectly benefit A.) In Pietraszewski's remaining conflict scenarios, A attacks C or vice versa, and we learn that λ_{AC} is negative and can thus infer that λ_{BC} is positive.

In each of Pietraszewski's four conflict scenarios, the two individuals he specifies as belonging to a group coincide with the two that can be inferred to share a mutually high welfare trade-off ratio according to the principles described above. A more parsimonious account, therefore, is to posit that mutually high welfare trade-off ratios can provide a foundation for recognizing social groups. Similar to Pietraszewski's proposal, this approach can be specified computationally; for instance, balance theory can be instantiated in the form of signed graphs in which the sign of each edge is the product of the signs of adjacent edges.

In contrast to Pietraszewski's proposal, a representation of social groups based on welfare trade-off ratios can generalize to situations without conflicts. For instance, A, B, and C could all have positive welfare trade-off ratios (under balance theory, a balanced triad can have 0 or 2 negative edges), resulting in mutually supportive behaviors and leading observers to consider all three a single social group. This proposal can, thus, better explain why prosocial and collective, interdependent behaviors can also serve as the basis for identifying groups and their members – for example, working together toward a shared goal (Sherif, 1966), or producing coordinated music or dance, which provides a credible signal of shared goals (Mehr, Krasnow, Bryant, & Hagen, 2021; Savage et al., 2021).

Mutually high welfare trade-off ratios are likely to be a strong cue to social group composition across many contexts, and often sufficient to define a social group on their own. They may also be developmentally privileged, serving as the core of infants' and young children's concepts of social affiliation (Noyes & Dunham, 2020; Powell, 2021). But, despite having more range than Pietraszewski's conflict-based primitives, welfare trade-off ratios still seem insufficient to capture the full range of social groups created and recognized by human adults.

Adult social groups vary widely in both size – from pairs to entire nations – and in permanence – from strangers who coordinate a onetime "flash mob," to religious groups that persist for millennia. Allowing welfare trade-off ratio calculations to be context-specific helps capture some of this diversity, but not all. For example, the United States Congress could be considered a social group, yet its members can feel such mutual animosity that they are inspired to political, and sometimes even physical, attacks.

To ask for a single computational primitive, or even a set of primitives, that can capture the vast array of social structures may not be reasonable. Instead, we propose that there are likely to be multiple computational strategies by which adults recognize and reason about social groups, with many features sufficient for identifying a group, but none strictly necessary. This does not merely refer to the capacity to learn statistical associations between surface characteristics such as clothing or race, and underlying coalitions (Kurzban, Tooby, & Cosmides, 2001), or to discern latent groupings of welfare trade-off ratios (Lau, Pouncy, Gershman, & Cikara, 2018). The ways in which humans use norms and rules to create groups and institutions of many varied forms require conceptual resources with more structure. Intuitive causal theories of social relationships and conventions, as well as our ability to analogize new groups to past social structures we have experienced, allow us to reason about disparate groups in disparate ways depending on our theories of their origins and characteristics (e.g., essentialized social groups, Hirschfeld, 1996; Rhodes, 2013; and institutional social groups, Noyes & Dunham, 2020).

Nonetheless, we are sympathetic to Pietraszewski's assertion that there is something conceptually primitive about the underlying relations captured by his conflict scenarios, which we argue are best instantiated as dyadic welfare trade-off ratios. Perhaps, it could be said that sets of mutually high welfare trade-off ratios provide the prototype for our concept of a social group: A collection of people willing to work toward individual or collective aims in the face of any nature of challenge.

Financial support. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review, 63*(5), 277–293.
Delton, A. W., & Robertson, T. E. (2016). How the mind makes welfare tradeoffs: Evolution, computation, and emotion. *Current Opinion in Psychology, 7,* 12–16.
Hirschfeld, L. A. (1996). *Race in the making: Cognition, culture, and the child's construction of human kinds.* The MIT Press.
Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences, 98*(26), 15387–15392.
Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General, 147*(12), 1881–1891.
Mehr, S. A., Krasnow, M. M., Bryant, G. A., & Hagen, E. H. (2021). Origins of music in credible signaling. *Behavioral and Brain Sciences, 44*, e60: 23–39.
Noyes, A., & Dunham, Y. (2020). Groups as institutions: The use of constitutive rules to attribute group membership. *Cognition, 196*, 104143.
Powell, L. J. (2021). Adopted utility calculus: Origins of a concept of social affiliation. *PsyArxiv.* Available at: https://doi.org/10.31234/osf.io/kuwgf.
Rhodes, M. (2013). How two intuitive theories shape the development of social categorization. *Child Development Perspectives, 7*(1), 12–16.
Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Music as a coevolved system for social bonding. *Behavioral and Brain Sciences, 44,* e59: 1–22.
Sherif, M. (1966). *In common predicament: Social psychology of intergroup conflict and cooperation.* Houghton Mifflin.
Tooby, J., & Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In *Handbook of emotions* (3rd ed, pp. 114–137). The Guilford Press.

Advantages and limitations of representing groups in terms of recursive utilities

Setayesh Radkani, Ashley J. Thomas and Rebecca Saxe

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA radkani@mit.edu; ajthomas@mit.edu; saxe@mit.edu, saxelab.mit.edu/ | doi:10.1017/S0140525X21001394, e118

Abstract

Group representations based on recursive utilities can be used to derive the same predictions as Pietraszewski in conflict situations. Additionally, these representations generalize to non-conflict situations, asymmetric relationships, and represent the stakes in a conflict. However, both proposals fail to represent asymmetries of power and responsibility and to account for generalizations from specific observed individuals to collections of non-observed individuals.

Pietraszewski argues that the representation of a group is defined by four ways a third person can be drawn into a dyadic conflict. We agree that such triadic interactions can lead to group representations, but propose that representations of groups are defined in terms of an abstract, recursive utility calculus (Kleiman-Weiner, Saxe, & Tenenbaum, 2017; Powell, 2021). By recursive utility we mean: People represent individuals as valuing (i.e., adopting, or weighting) the utilities of other individuals.

Representing recursive utilities licenses the same inferences as the triadic behavioural primitives, in the conflicts Pietraszewski considered. If A and C put high weight on each other's utilities, but relatively lower or even negative weight on B's utilities, then observers are licensed to predict: C will attack B, defend A, not be attacked by A, and possibly be attacked by B. Observing one of these patterns allows an observer to infer the relative weights these individuals place on each other's utilities, and therefore to predict their future behaviour.

However, using recursive utilities allows the theory to extend beyond Pietraszewski's account in three ways.

First, the recursive utility representation accounts for how people represent and learn about groups in non-conflict situations. For example, given the pattern of utilities described above observers can predict that C is more likely to respond to A's needs than to B's needs, outside of conflict. Observing C helping or caring for A would provide some evidence of how much C values A's utilities (Powell, 2021). Then, at the first sign of a conflict between A and B, an observer would predict C's role, even though no previous conflict-related behaviour had ever been observed.

Second, the recursive utility representation accounts for how people represent and learn about asymmetric relationships. The inferential links between Pietraszewski's primitives are symmetric: When C is drawn into the conflict between A and B, and attacks B, all of three remaining primitives are predicted to the same degree. However, C attacking B provides more evidence of how much C values A relative to B, than about how much A values C. The weights individuals put on each others' utilities do not need to be symmetric (Powell, 2021). Recursive utilities allow observers to represent cases where C would always ally with A against B, but A would not reciprocate.

Third, representing conflicts in terms of abstract utilities could account for how people make predictions based on what the conflict is about. Two individuals may become allies when parts of their utilities overlap, and they act to maximize joint utilities (Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016). Unlike fully adopting another individual's utilities, alliances built on joint utilities might be limited to specific conflicts. For example, hunters and environmentalists may find common cause in defending public access to wildlands but be opposed on animal rights. To predict future behaviour, it is critical to identify not only who is fighting, but also what is at stake.

A computational theory of groups in terms of abstract, recursive utilities is thus appealing for its expressive range and inferential flexibility. Note that implementing an actual computational model with these properties is very challenging. No existing computational model distinguishes and/or combines both recursive and joint utilities and their interactions.

Yet this proposal shares key limitations with Pietraszewski's.

Both Pietraszewski's behavioural primitives, and the recursive utility representation, represent triads of individuals who have equal power and responsibility. To predict human behaviour in conflict situations, though, asymmetries of power and responsibility are indispensable. For example, if a parent defends their small child from attack, we do not predict that the child is equally likely to defend their parent (because of differences in responsibility) (e.g., Fiske, 1992); when a person is insulted by his boss, rather than his co-worker, we can predict that fewer observers will come to his defense (because of differential power). These asymmetries cannot be reduced to (or derived from) triads of alliance or recursive value, and so require distinct cognitive machinery.

The other major challenge is how people generalize group-constitutive roles from specific observed individuals to collections of unobserved individuals. Pietraszewski calls this "substitution": commonly, the parties to a conflict are not literal individuals, but potentially large sets of individuals who are substitutable for one another in the conflict roles. But how do observers identify these sets? In the recursive utility framework, the same problem arises: How do observers infer the recursive utilities of new individuals? The intuitive answer here is: Observers use observable ("ancillary") cues and culturally specific intuitive theories (O'Connor, 2019) to guess whose utilities are substitutable. These inferences may be well-founded (e.g., in an international war, most individuals will have higher reciprocal weights on the utilities of people who share their nationality, than people from the opposing nationality), but may be wrong (e.g., in every war, some people sympathize with the individual or collective good of people from the "other" nationality); and because they are always based on limited evidence, these inferences are often stereotypes (e.g., people who share a racial or gender or religious identity may be viewed as "substitutable" in absurd and offensive ways) (e.g., Bruneau, Kteily, & Falk, 2018). Neither account explains the role of shared historical knowledge or cultural learning in these inferences.

To determine who is substitutable in a conflict role, observers could generalize across people based on theories and observations of social organization. Yet this solution seems to require all of the machinery of group membership that Pietraszewski claimed to avoid. To infer which individuals have relevantly similar utilities, or are substitutable in conflict roles, observers need to generalize between people based on dyadic shared properties, rather than triadic conflict interactions.

In summary, a representation of abstract, recursive utilities could make the same predictions as Pietraszewski, and additionally account for how people represent and learn about groups in non-conflict interactions, represent asymmetric relationships within groups, and make predictions based on what the conflict is about. However, two key limitations of Pietraszewski's approach are shared by our proposal, and thus remain to be addressed.

Financial support. SR and RS were supported by Patrick J. McGovern Foundation Grant. AJT was supported by NIH National Research Service Award 1F32HD096829.

Conflict of interest. None.

References

Bruneau, E., Kteily, N., & Falk, E. (2018). Interventions highlighting hypocrisy reduce collective blame of Muslims for individual acts of violence and assuage anti-Muslim hostility. *Personality and Social Psychology Bulletin, 44*(3), 430–448.
Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review, 99*(4), 689.
Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.
Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition, 167*, 107–123.
O'Connor, C. (2019). *The origins of unfairness: Social categories and cultural evolution*. Oxford University Press, USA.
Powell, L. (2021). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916211048487

On vagueness and parochialism in psychological research on groups

Kyle G. Ratner[a], David L. Hamilton[a] and Marilynn B. Brewer[b]

[a]Department of Psychological and Brain Sciences, University of California,
Santa Barbara, Santa Barbara, CA 93106, USA and [b]Department of Psychology, The Ohio State University, Columbus, OH 43210, USA
kyle.ratner@psych.ucsb.edu | https://spl.psych.ucsb.edu | doi:10.1017/S0140525X21001369, e119

Abstract

Pietraszewski asserts that social psychological research on groups is too vague, tautological, and dependent on intuitions to be theoretically useful. We disagree. Pietraszewski's contribution is thought-provoking but also incomplete and guilty of many of the faults he attributes to others. Instead of rototilling the existing knowledge landscape, we urge for more integration of new and old ideas.

Pietraszewski's analysis of "what is a group?" is novel and stimulating. However, the gusto with which he dismissed the existing research as vague was overwrought and counterproductive. Except for an occasional citation with minimal elaboration, he did not engage with the social psychological work he dismisses and he omits several relevant contributions. We particularly found this puzzling because it is not obvious to us that Pietraszewski's theorizing provides any more clarity and specificity than the research that he criticizes.

Social psychologists interested in group dynamics have worked on tightening their conceptual understanding of psychological representation and process by adapting mental models from cognitive science. The concept and category literature was brought to bear on how people represent themselves and others as individuals and as group members (e.g., Brewer, 1988; Fiske & Neuberg, 1990) and the formation of stereotypes and prejudice were interpreted through models of attention and memory (e.g., Devine, 1989; Greenwald & Banaji, 1995; Hamilton & Gifford, 1976). As cognitive science (and related motivation, affective, and neural sciences) advanced, social psychologists updated their definitions accordingly (e.g., Amodio and Ratner, 2011). Although more research is undoubtedly needed, the social psychology of groups is not the straw man that Pietraszewski depicts.

Moreover, the standard that Pietraszewski uses when evaluating the research of others is not the one he uses for himself. He criticizes social psychologists for relying too strongly on intuitions, but his theorizing seems very much based on his own intuitions. Pietraszewski provided no empirical evidence to support his assertions. It is unclear why, for instance, he assumes that perceivers inherently view the behaviors in his primitives as evidence for intergroup behavior instead of a string of dyadic interpersonal behaviors. In Figure 3 he circumvents this ambiguity by labeling some positions in the diagram as ingroup and some as outgroup. However, this solution is as tautological as the container metaphor he chastises. He also uses rhetorical sleight of hand to excuse vagueness in his own theorizing. He limits the scope of his analysis to intergroup conflict when it is convenient for him, although his aims often seem broader. This scope-narrowing for himself is particularly glaring because he criticizes the field as a whole for not being able to account for how groups, in general, are represented. He also frames his ideas as simply working toward a comprehensive theory, which allows him to defer the hard work to future directions. He ignores that sometimes people derogate ingroup members more than outgroup members (e.g., Marques, Yzerbyt, & Leyens, 1988). He leaves vague how people reason about groups when they are not privy to observing behaviors of people over time and how his reasoning applies to representing groups that are not in conflict. He also derides similarity-based definitions of groups, but he seems to rely on perceivers inferring similar fate and goals of allied agents when analyzing the group-constitute roles.

By dismissing social psychology he misses an opportunity to engage with research that is relevant to his research. We offer two examples. First, his interest in understanding the nature of groups from an evolutionary perspective reminded us of Caporael's core configurations model (1997; Brewer & Caporael, 2006). Caporael's analysis discusses different kinds of groups (dyad, work/family group, deme, and macrodeme), the functions of these groups, and the tasks that each size group facilitates. Group size and the different functions of groups are not taken into account by Pietraszewski and considering these variables in relation to Caporael's research could be useful for thinking about how Pietraszewski's primitives might scale and what other primitives could be considered.

Second, other social psychology research, based on data-driven analyses, has shown that groups can be clustered into five different types: intimacy groups (e.g., families), task-oriented groups (e.g., work team), social categories (e.g., those defined on the basis of gender, race, nationality), loosely associated groups (e.g., individuals employed by the same large technology company), and transitory groups (e.g., people waiting at a particular airport terminal) (Lickel et al., 2000). These groups differ in their perceived entitativity, in the pattern of features used to determine that a collection of people constitute a group, and in how information about the groups are stored in memory. A series of experiments (Sherman, Castelli, & Hamilton, 2002) demonstrated that people not only understand these group type distinctions, but they also spontaneously use them in processing and storing information. In addition, Johnson et al. (2006) address the question of function, showing that these different group types serve different human needs or motives (affiliation, achievement, and identity). It is also notable that the way that Pietraszewski talks about ancillary attributes is reminiscent of Hamilton, Sherman, & Lickel's (1998) application of Brunswick's lens model to understand entitativity. For Pietraszewski to achieve his larger aims of creating a grand computational theory about groups he would need to consider many of the issues that have been discussed by these researchers.

From our vantage, Pietraszewski's theorizing does not supplant existing research and is instead on the same theoretical plane. We believe that theoretical integration rather than competition between models of "groupness" is the best path forward. We laude Pietraszewski's call for more theoretical specificity and see it as consistent with recent pleas by mathematical psychologists (e.g., Guest and Martin, 2021) calling for psychologists across all topic areas to more explicitly declare their theoretical premises. We further appreciate how Pietraszewski invokes Marr's levels of understanding to illustrate the complementary ways that an information processing theory of groups should be articulated. Although psychologists interested in social dynamics have found Marr's organization useful (e.g., Cunningham, Zelazo, Packer, & Van Bavel, 2007; Lockwood, Apps, & Chang, 2020; Mitchell, 2006), the literature on groups (to our knowledge) had not previously taken advantage of this framework. We are optimistic that greater integration across psychology's subdisciplines and adjacent fields coupled with recent efforts to embrace open science, to use more naturalistic experimental paradigms, and to recognize that cognition can be distributed across individuals, will continue to improve definitional specificity and validity in the psychological investigation of group living.

Financial support. The writing of this commentary was not supported by a specific grant from any funding entity.

Conflict of interest. None.

References

Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science, 20*(3), 143–148.
Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull, & R. S. Wyer, Jr. (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Erlbaum.
Brewer, M. B., & Caporael, L. R. (2006). An evolutionary perspective on social identity: Revisiting groups. In M. Schaller, J. Simpson, & D. Kenrick (Eds.), *Evolution and social psychology* (pp. 143–161). Psychology Press.
Caporael, L. R. (1997). The evolution of truly social cognition: The core configurations model. *Personality and Social Psychology Review, 1*(4), 276–298.
Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition, 25*(5), 736–760.
Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18.
Fiske, S. T., & Neuberg, S. L. (1990). A continuum model of impression formation, from category-based to individuating processes: Influence of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Academic Press.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27.
Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science, 16*(4), 789–802.
Hamilton, D., & Gifford, R. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology. 12*(4), 392–407.
Hamilton, D. L., Sherman, S. J., & Lickel, B. (1998). Perceiving social groups: The importance of the entitativity continuum. In C. Sedikides, J. Schopler, & C. A. Insko (Eds.), *Intergroup cognition and intergroup behavior* (pp. 47–74). Erlbaum.
Johnson, A. L., Crawford, M. T., Sherman, S. J., Rutchick, A. M., Hamilton, D. L., Ferreira, M. B., & Petrocelli, J. V. (2006). A functional perspective on group memberships: Differential need fulfillment in a group typology. *Journal of Experimental Social Psychology, 42*(6), 707–719.
Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A. C., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology, 78*, 223–246.
Lockwood, P. L., Apps, M. A., & Chang, S. W. (2020). Is there a "social" brain? Implementations and algorithms. *Trends in Cognitive Sciences, 24*(10), 802–813.
Marques, J. M., Yzerbyt, V. Y., & Leyens, J.-P. (1988). The black sheep effect: Extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology, 18,* 1–6.
Mitchell, J. P. (2006). Mentalizing and Marr: An information processing approach to the study of social cognition. *Brain Research, 1079*(1), 66–75.
Sherman, S. J., Castelli, L., & Hamilton, D. L. (2002). The spontaneous use of a group typology as an organizing principle in memory. *Journal of Personality and Social Psychology, 82*(3), 328.

Towards a computational network theory of social groups

Daniel Redhead, Riana Minocher and Dominik Deffner

Department of Human Behaviour, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany daniel redhead@eva.mpg.de | doi:10.1017/S0140525X21001370, e120

Abstract

Network theory is necessary for the realization of cognitive representations and resulting empirical observations of social groups. We propose that the triadic primitives denoting individual roles are multilayer, with positive and negative relations feeding into cost–benefit calculations. Through this, we advance a computational theory that generalizes to different scales and to contexts where conflict is not present.

Pietraszewski advances a compelling computational theory of how the mind represents social groups. Here, we incorporate several key advances in network science and social network analysis – fields built upon a core set of theory and analytical tools for understanding the structure of social systems. In doing this, we propose an extension and reframing of Pietraszewski's computational theory of social groups that provides intuitive building blocks for generative models that (a) translate to observable patterning of social groups in real-world settings, (b) do not rely on conflict dynamics, and (c) easily scale to characterize higher-order group settings.

In line with core theory in sociology (Blau, 1964; Homans, 1958; Simmel, 1950), within Pietraszewski's framework any social behaviour is considered a transaction or exchange, with individuals acting based on the net rewards of such relations (Caplow, 1956; Emerson, 1976). In particular, these notions reflect key components of network theory, where the cost–benefit calculations of forming a social relationship may be weighted by the structure and dynamics of existing relationships (e.g., reciprocity or triadic configurations; Cropanzano & Mitchell, 2005; Doreian & Krackhardt, 2001). The probability that individuals occupy a given role within the triad – for example, that they engage in conflict against a lone other – is likely determined by observable individual differences, or ancillary attributes. These ancillary attributes may be socially constructed categorical types that delineate group membership (such as race or ethnicity; Pietraszewski, Cosmides, & Tooby, 2014). Quantitative individual differences are equally likely to enter into the event calculus, with individuals' power, status, or resource-holding potential feeding into choices about engaging in conflict (or any other social behaviour; Redhead, Dhaliwal, & Cheng, 2021; von Rueden, Redhead, O'Gorman, Kaplan, & Gurven, 2019).

Incorporating network theory, we reframe Pietraszewski's account as an instance of mixed "triadic closure" in "signed networks" (i.e., social networks that have both positive and negative relationships; Cartwright & Harary, 1956; Heider, 1958). A third, positive tie must be specified between two individuals to develop realistic generative models that reliably determine individual roles within any triadic configuration. This third tie could represent any number of social relationships or behaviours (such as kinship, friendship, physical proximity, food-sharing, or co-working) that enter the cost–benefit analysis. In the context of conflict, the tie is assumed to be "positive," producing benefits or rewards within the "event calculus." Conflicts are then assumed to be "negative" ties (i.e., that confer action costs). Complementary to Pietraszewski's formulation, we can consider a group as being a dyad or larger set of individuals connected by observable benefit-generating relationships.

Social groups may, therefore, be defined and represented in a way that does not rely on conflict dynamics. This logic can be applied to determine acts that generate benefits, with individuals transferring portions of a finite set of resources (e.g., food, labour, or money) based on the different token or event types within a triad. This may produce an extended number of triadic configurations – either open or closed – that could arguably be different summary representations of social groups (Block, 2015).

Including the content of a third tie provides concrete and flexible representations of social groups. This is because group membership – per Pietraszewski's definition – is fundamentally a property of a dyad. That is, within a given triad, two individuals must share a group membership. Our marking of group membership, therefore, does not solely rely on there being observable individual markers, but also considers the presence of social relationships. This framework can more easily be extended for group membership to be dynamic (i.e., not strictly fixed), and for individuals to operate within multiple group memberships. Adding in this further complexity may be fruitful, as individuals operate in complex, overlapping relationships (i.e., multilayer networks; Kivelä et al., 2014), and these relationships may also enter any cost–benefit calculations, allowing easy translation to observed empirical patterns. That is, in many real-world settings, individuals may face a decision to engage in a conflict between the members of two different social groups that they are members of.

We can then extend the model to incorporate information about these different group memberships into the event calculus that may, for example, establish rules about roles and representations of groups when conflicts arise between kin, friends, or cooperative partners (Redhead & von Rueden, 2021).

Bringing this all together, our reformulation of Pietraszewski's theory provides a more principled approach to group settings that extend beyond the triad and can easily scale to bounded groups (or networks) of any size. If groups are not observed but ties are, we can calculate probabilities of individuals' roles within a triad, and these roles and relationships will aggregate up to modular population-level networks that will comprise two or more "communities" (e.g., social groups) that we can analytically detect (e.g., Amelio & Pizzuti, 2016). That is, the information presented about ties may be used to determine group structure across any scale, and has been applied to large-scale international military data to determine groups of countries that form alliances and conflicts (Traag & Bruggeman, 2009). If communities are "observed" – as in the example in Figure 3 given by Pietraszewski – this process becomes somewhat analogous to a class of generative network models (stochastic block models for signed networks) that use observed group structure to determine and/or predict tie formation within and between groups (e.g., Doreian & Mrvar, 2009). In sum, by reframing and extending Pietraszewski's account to incorporate network theory, we are able to establish summary representations of social groups that do not rely on conflict, easily scale to group settings that go beyond the triad, and provide an architecture for generative models of social groups. Through this, a network approach provides a framework for linking abstract cognitive concepts to empirically observed patterns in real-world settings.

Financial support. Daniel Redhead, Riana Minocher, and Dominik Deffner were supported by the Department of Human Behaviour, Ecology and Culture at the Max Planck Institute for Evolutionary Anthropology.

Conflict of interest. The authors have no conflicts of interest to declare.

References

Amelio, A., & Pizzuti, C. (2016). An evolutionary and local refinement approach for community detection in signed networks. *International Journal on Artificial Intelligence Tools, 25*(04), 1650021.
Blau, P. M. (1964). Justice in social exchange. *Sociological Inquiry, 34*(2), 193–206.
Block, P. (2015). Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks, 40,* 163–173.
Caplow, T. (1956). A theory of coalitions in the triad. *American Sociological Review, 21*(4), 489–493.
Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review, 63*(5), 277.
Cropanzano, R., & Mitchell, M. S. (2005). Social exchange theory: An interdisciplinary review. *Journal of Management, 31*(6), 874–900.
Doreian, P., & Krackhardt, D. (2001). Pre-transitive balance mechanisms for signed networks. *Journal of Mathematical Sociology, 25*(1), 43–67.
Doreian, P., & Mrvar, A. (2009). Partitioning signed social networks. *Social Networks, 31*(1), 1–11.
Emerson, R. M. (1976). Social exchange theory. *Annual Review of Sociology, 2,* 335–362.
Heider, F. (1958). *The psychology of interpersonal relations*. Psychology Press.
Homans, G. C. (1958). Social behavior as exchange. *American Journal of Sociology, 63*(6), 597–606.
Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks, 2*(3), 203–271.
Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. *PLoS One, 9*(2), e88534.
Redhead, D., Dhaliwal, N., & Cheng, J. T. (2021). Taking charge and stepping in: Individuals who punish are rewarded with prestige and dominance. *Social and Personality Psychology Compass, 15*(2), e12581.

Redhead, D., & von Rueden, C. R. (2021). Coalitions and conflict: A longitudinal analysis of men's politics. *Evolutionary Human Sciences, 3,* E31.
Simmel, G. (1950). *The sociology of Georg Simmel* (Vol. 92892). Free Press, Macmillan Publishers.
Traag, V. A., & Bruggeman, J. (2009). Community detection in networks with positive and negative links. *Physical Review E*, 80(3), 036115.
von Rueden, C. R., Redhead, D., O'Gorman, R., Kaplan, H., & Gurven, M. (2019). The dynamics of men's cooperation and social status in a small-scale society. *Proceedings of the Royal Society B, 286*(1908), 20191367.

Social groups and the computational conundrums of delays, proximity, and loyalty

Dragos Simandan

Geography Department, Brock University, St. Catharines, ON L2S 3A1, Canada simandan@brocku.ca | https://brocku.ca/social-sciences/geography/faculty-and-staff/dragos-simandan/ | doi:10.1017/S0140525X21001205, e121

Abstract

Even though Pietraszewski acknowledges the tentative nature of the theory and the multiple lines of adjacent research needed to flesh it out, he insists that the finite set of primitives he identified is necessary and sufficient for defining social groups in the context of conflict. In this commentary I expose three interrelated conundrums that cast doubt on this simplistic presumption.

David Pietraszewski's proposal for a computational theory of social groups is a significant original contribution to psychology, cognitive science, and the social sciences, as it offers a nontautological and relational understanding of social groups by means of a finite collection of cognitive primitives. What I find most commendable in the target article is the attempt to develop a relational account of social groups without compromising the epistemic virtues of precision and clarity: for context, in my home discipline of human geography, relational understandings of social and spatial phenomena too often have come at the high price of imprecise, unclear, and metaphor-laden jargon (cf. Malpas, 2012). Interestingly, Pietraszewski's incisive critique of the still-dominant theorizing of groups as containers in which agents are somehow placed (cf. sect. 8.2, paras. 2–4) and the proposed redefinition of group membership as a relational property parallels recent critiques in human geography of (social) space itself as container and its rethinking in relational terms (Simandan, 2019a, 2020).

Even though the author acknowledges the tentative nature of the theory and the multiple lines of adjacent research needed to flesh it out, he insists without much warrant that the finite set of primitives he identified in sections 3 (paras. 3–4) and 4 (paras. 2–3) is necessary and sufficient for defining social groups in the context of conflict. In the remainder of this commentary, I expose three interrelated conundrums that undermine this simplistic presumption.

The first is the underappreciated role of delays in the computational problem of representing and reasoning about groups. Although the author briefly mentions that "contingent cost-infliction is often a drawn-out process, with many gaps and lulls between interactions" (sect. 6, para. 3) and that "therefore, delays between cost inflictions will have to be allowed for" (sect. 6, para. 3), he fails to develop the logical implications of this observation for how we make sense of groups-in-conflict. To appreciate this point, we need a more elaborate vocabulary for conceptualizing delays in the context of intergroup conflict (cf. Simandan, 2018, 2019b, 2019c). To begin with, a group's response to a particular challenge or move by a rival group can be immediate or delayed. This basic distinction is complicated by the fact that even apparently immediate reactions always involve a number of unavoidable delays pertaining to observation (delay between the emergence of a relevant signal in the environment and one's taking notice of that information), initial insight or *coup d'oeil* (delay between being aware of disparate signals and their subsequent juxtaposition into an incipient mental schema), full mental model development (delay between initial inchoate insight and its gradual development into a full representation of the group conflict situation), intragroup communication or reporting delays, group decision-making delays, initiation or launching (delay between a group's decision to respond to a rival group's move and the actual moment when that response is carried out), and material delays (the irreducible gap between launching a given response and the time it takes for it to generate results). The formal analysis of the dynamics of intergroup conflict requires dedicated attention to teasing apart unavoidable delays from premeditated or deliberate delays in move-countermove sequences. It also requires a more explicit acknowledgment that delays are a fundamental category that needs to be accounted for in any attempt to formalize the meaning of social groups beyond natural language understanding. This latter point unintentionally transpires even from Pietraszewski's proposed formalism (sects. 3 and 4): Indeed, although he builds his definition of groups-in-conflict by relying on the four triadic primitives of generalization ("A attacks B, then A attacks C"), alliance ("A attacks B, then C attacks B"), displacement ("A attacks B, then B attacks C"), and defense ("A attacks B, then C attacks A"), one of the common threads linking these triadic primitives is that they all function with delays, as signaled by the marker "then" (my emphasis).

The second conundrum results from Pietraszewski's glib relegation of proximity to the status of a mere ancillary attribute of social groups (cf. sect. 8.1, para. 2, and again in sect. 9.2, para. 2). I argue that once we take into account the multidimensional understandings of proximity/distance developed in construal-level theory and evolutionary geography, it becomes apparent that proximity is a strong candidate for inclusion in the set of primitives that defines what a social group is (cf. Simandan, 2016). Construal-level theory (Trope & Liberman, 2010) has advanced a subjective account of proximity whereby distance is a metric that tracks how far removed from the self-in-the-here-and-now an item is alongside the four interrelated dimensions of space, time, uncertainty, and sociality. Evolutionary geography (Boschma, 2005) has also moved toward a multidimensional conceptualization of proximity, identifying no less than five types: cognitive (degree of overlap between the dominant mental representations of two groups or organizations), geographical (physical objective distance), organizational (extent of shared history or past cooperation between two groups), institutional (degree of overlap between the cultures of two groups), and social proximity (intensity of social and kinship ties between two agents or groups). Both of the aforementioned theories identify social proximity as one of the crucial dimensions of proximity. In natural language, we use the vocabularies of social proximity and social groups seamlessly and often interchangeably: I would therefore urge Pietraszewski to develop his formal account of groups by exploring ways to upgrade proximity (understood multidimensionally) from ancillary to central status.

Finally, Pietraszewski's proposed building blocks for a computational account of social groups reveal repeatedly that the problematic of loyalty and disloyalty seems to be inextricably intertwined with how agents make inferences about groups (e.g., sect. 8.2, para. 6 and sect. 8.3, para. 2; sect. 8.4, para. 3). Which brings out the question: What would a computational theory of group loyalty itself look like? The philosophical literature on loyalty may not offer useful starting points because it often frames this topic as "an important area of the normative" (Oldenquist, 1982, p. 173). A more promising starting point seems to be appraising the dynamic relationship between social proximity and loyalty, and thereby pushing the *relational* understanding of social groups one step further.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors

Conflict of interest. None.

References

Boschma, R. (2005). Proximity and innovation: A critical assessment. *Regional Studies, 39*(1), 61–74.
Malpas, J. (2012). Putting space in place: Philosophical topography and relational geography. *Environment and Planning D: Society and Space, 30*(2), 226–242.
Oldenquist, A. (1982). Loyalties. *The Journal of Philosophy, 79*(4), 173–193.
Simandan, D. (2016). Proximity, subjectivity, and space: Rethinking distance in human geography. *Geoforum, 75,* 249–252.
Simandan, D. (2018). Competition, contingency, and destabilization in urban assemblages and actor-networks. *Urban Geography, 39*(5), 655–666.
Simandan, D. (2019a). Revisiting positionality and the thesis of situated knowledge. *Dialogues in Human Geography, 9*(2), 129–149.
Simandan, D. (2019b). Competition, delays, and coevolution in markets and politics. *Geoforum, 98,* 15–24.
Simandan, D. (2019c). Iterative lagged asymmetric responses in strategic management and long-range planning. *Time & Society, 28*(4), 1363–1381.
Simandan, D. (2020). Being surprised and surprising ourselves: A geography of personal and social change. *Progress in Human Geography, 44*(1), 99–118.
Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463.

Shadow banning, astroturfing, catfishing, and other online conflicts where beliefs about group membership diverge

Jordan W. Suchow

School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA jws@stevens.edu; http://suchow.io |
doi:10.1017/S0140525X21001448, e122

Abstract

Drawing from conflicts observed in online communities (e.g., astroturfing and shadow banning), I extend Pietraszewski's theory to accommodate phenomena dependent on the intersubjectivity of groups, where representations of group membership (or beliefs about group membership) diverge. Doing so requires enriching representations to include other agents and their beliefs in a process of recursive mentalizing.

In the target article, Pietraszewski proposes a computational theory of social groups that is at its core *subjective*, defining a group in terms of a single individual's representation of it. However, social groups are not subjective: Consider that a person cannot through their own beliefs unilaterally create or destroy a group, or change an established group's membership. Rather, group membership is *intersubjective*, dependent on representations that are (at least in part) shared among members of the ingroup or outgroup (Dennen & Wieland, 2007; Eden, Jones, Sims, & Smithin, 1981; Matusov, 1996; Stahl, 2016; Zlatev, 2014). The intersubjective nature of groups gives rise to important phenomena in the context of conflict that cannot be explained by Pietraszewski's computational theory because they arise only when people's representations of group membership diverge or are believed to have diverged.

Consider the following examples often observed in conflicts within online communities:

(1) A person wrongly believes they are part of a group, whose members keep up the charade until an ultimate act of (apparent) betrayal that reveals the false belief (e.g., cyberbullying).
(2) A person becomes romantically involved with a defrauder, scammer, troll, or person with some other ulterior motive (e.g., online dating romance scams; catfishing).
(3) A person undermines a group by pretending to be a member of it while covertly acting against its interests (e.g., online sock puppetry and astroturfing; cyber espionage).
(4) Two factions of a group each reject the others' sincerely held beliefs regarding membership in that group, cleaving it in two (e.g., schisms).
(5) An aggressor creates a self-fulfilling prophecy when, failing to distinguish between groups, aggresses against them jointly, causing the groups to merge (e.g., in the emergence of some unity movements).
(6) One person does not believe a particular group exists, whereas another person cherishes their membership in that group (e.g., identity pride and erasure).
(7) A person is unaware of having been banished from a group or silenced within it (e.g., hell banning and shadow banning).

Analyzing any of these phenomena by considering only one individual's representations of group membership would fail to capture their essence. For example, in the case of astroturfing (Leiser, 2016; Sisson, 2017; Zhang, Carpenter, & Ko, 2013), where a purported grass-roots organizer is, in fact, the agent of a sponsor working against the cause, it is not enough for the group or the public to believe the agent is a member of the group. Nor is it enough for the sponsor or its agent to believe they are not a member. Rather, it is the divergence in understanding about the agent across the sponsor, the agent, the public, and other interested parties, which gives rise to the phenomenon and the harmful consequences to the cause that are associated with it.

Here, I put forward a computational approach that extends Pietraszewski's theory to accommodate phenomena dependent on the intersubjectivity of groups.

The extension begins by zooming out from the focal individual studied by Pietraszewski to consider the representations of all three individuals in the triad (or more generally, any interested parties). Minimally, this is accomplished by endowing each individual with their own representation of the kind put forward by Pietraszewski in terms of which agents will tend to fill the group-constitutive roles. Doing so requires no new computational machinery and permits analysis of diverse phenomena where these representations diverge. For example, in the case of shadow banning, a person is exiled from an online community without their knowledge by a moderator who causes the exiled person's communication to be invisible to other community members (Cole, 2018; waxpancake, 2009). It is common for the shadow-banned individual, the moderator, and other community members to each have their own understanding: The shadow-banned individual believes they are part of the group, the moderator believes they are not, and other members of the group may variously believe the individual is a member, a non-member, or does not even exist. A meaningful description of a group must, therefore, allow expression of divergent representations of group membership.

The extension proceeds by enriching the content of each individual's representation to include the representations of other individuals via a process of recursive mentalizing. Although in the previous step, we endowed each agent with a representation that included other agents filling (or not) group-constitutive roles, but not those other agents' beliefs, in the current step we recurse, enabling each agent to represent other agents' beliefs (Frith & Frith, 2005). At infinite recursion depth, this produces effects of common knowledge (de Freitas, Thomas, DeScioli, & Pinker, 2019; Platow, Foddy, Yamagishi, Lim, & Chow, 2012; Thomas et al., 2016). New computational machinery in the form of recursive mentalizing must be brought to the table, bringing with it the power to model complex social phenomena that depend on misrepresentation and deception, where actions are taken because they are expected to validate another person's wrongly held beliefs or cause them to misinterpret which agents fill group-constitutive roles. Returning to the example of astroturfing, consider that divergence in representations alone is not enough to fully capture its essence – being mistaken as a member of a group is not astroturfing. Rather, the agent must also take an action because they believe it will cause a certain impression in the minds of the public with respect to group membership. A representation of a group in the context of conflict must, therefore, enable individuals to represent the beliefs of others.

Enriching the representation put forward by Pietraszewski to include other agents and their beliefs in a process of recursive mentalizing permits analysis of complex social phenomena that arise from the intersubjectivity of groups.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Cole, S. (2018). Where did the concept of "shadow banning" come from? Motherboard: Tech by Vice. https://www.vice.com/en/article/a3q744/where-did-shadow-banningcome-from-trump-republicans-shadowbanned.
de Freitas, J., Thomas, K., DeScioli, P., & Pinker, S. (2019). Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences, 116*(28), 13751–13758.
Dennen, V. P., & Wieland, K. (2007). From interaction to intersubjectivity: Facilitating online group discourse processes. *Distance Education, 28*(3), 281–297.
Eden, C., Jones, S., Sims, D., & Smithin, T. (1981). The intersubjectivity of issues and issues of intersubjectivity. *Journal of Management Studies, 18*(1), 37–47.
Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology, 15*(17), R644–R645.
Leiser, M. (2016). AstroTurfing, "CyberTurfing" and other online persuasion campaigns. *European Journal of Law and Technology, 7*(1), 1–27.
Matusov, E. (1996). Intersubjectivity without agreement. *Mind, Culture, and Activity, 3*(1), 25–45.
Platow, M. J., Foddy, M., Yamagishi, T., Lim, L. I., & Chow, A. (2012). Two experimental tests of trust in in-group strangers: The moderating role of common knowledge of group membership. *European Journal of Social Psychology, 42*(1), 30–35.
Sisson, D. C. (2017). Inauthentic communication, organization-public relationships, and trust: A content analysis of online astroturfing news coverage. *Public Relations Review, 43*(4), 788–795.
Stahl, G. (2016). From intersubjectivity to group cognition. *Computer Supported Cooperative Work (CSCW), 25*(4), 355–384.
Thomas, K. A., De Freitas, J., DeScioli, P., & Pinker, S. (2016). Recursive mentalizing and common knowledge in the bystander effect. *Journal of Experimental Psychology: General, 145*(5), 621–629.
waxpancake. (2009). What was the first website to hide troll's activity to everyone but the troll himself? Ask MetaFilter. https://ask.metafilter.com/117775/.
Zhang, J., Carpenter, D., & Ko, M. (2013). Online astroturfing: A theoretical perspective. *AMCIS* 2013.
Zlatev, J. (2014). The co-evolution of human intersubjectivity, morality and language. In *The social origins of language* (pp. 249–266). Oxford University Press.

More than one way to skin a cat: Addressing the arbitration problem in developmental science

Denis Tatone

Department of Cognitive Science, Central European University, 1100 Vienna, Austria TatoneD@ceu.edu; denis.tatone@gmail.com |
doi:10.1017/S0140525X21001400, e123

Abstract

David Pietraszewski's theory of social groups offers a developmentally plausible account of how we reason about group membership, as it delineates clear boundaries to the hypothesis space that children must navigate. Merits notwithstanding, the account remains silent with respect to the arbitration problem: It does not explain how children can appropriately select among competing frames when interpreting social interactions.

From a developmental standpoint, the main virtue of David Pietraszewski's theory is its ability to deal effectively with the reduction problem. At its core is the claim that the gamut of multi-agent conflict can be decomposed into four types of triadic interactions, each specifying a distinct way by which third parties can be drawn into conflict. By constraining the hypothesis space to a finite repertoire of coalitional schemata, the theory offers a developmentally plausible way in which young learners may infer role assignments across multi-agent configurations. Once the appropriate stance is adopted, attributing group membership becomes an eminently tractable task (cf. Thomsen & Carey, 2013).

Nothing, however, guarantees that third-party interactions end up being interpreted through coalitional lenses. Any given social behavior can be, in fact, apprehended under multiple and equally compatible frames. By way of example, let us consider the "generalization" event, in which A imposes costs on B and C. By Pietraszewski's account, this event should license the inference that B and C belong to the same group. This is the case, however, only if the observer is trying to map the observed interactions in terms of coalitional units. Other interpretations are certainly possible. For instance, the observer may take the across-patient consistency of the agent's action as evidence of an underlying trait (i.e., A is a bully; Boseovski & Lee, 2006). By focusing on the attribution of individual dispositions, this stance inhibits utilizing B and C's common interaction role (as victims) as group-diagnostic information. Reasoning about traits is, of course, not the only contender. Triadic interactions may be revelatory of underlying social structures. Let us consider the "displacement" case: A imposes costs on B, who does the same on C. Under Pietraszewski's account, A and C should be represented as belonging to the same group. This interaction can, however, also be taken to indicate a linear dominance structure in which a high-ranking agent (A) attacks a subordinate (B), who redirects its aggression against a lower-ranking agent (C) to prevent further aggression from third parties (for evidence of this behavior in nonhuman animals, see Kazem & Aureli, 2005). This interpretation, although similarly requiring the embedding of pairwise relations within a larger structural template, does not license the assumption that A and C are in the same group.

As these examples illustrates, there are multiple frames to choose among. This plurality is to be expected, as it reflects the variety of fitness-relevant affordances in the social world. If the interpretation of social interactions can be considered akin to the phenomenon of multistable perception, systems for arbitrating among these competing frames then must be expected, especially when the concurrent activation of multiple frames may lead to contradictory inferences (as in the "displacement" scenario, in which two agents would be represented at the same time as being part of the same group as well as placed at the opposite ends of a hierarchy). If Pietraszewski's computational theory manages to effectively imprint deterministic force to group-based reasoning in a specific coalitional frame, it offers little guidance with respect to these broader arbitration issues. This architectural blind spot is, however, orthogonal to the internal validity of Pietraszewski's theory: Once a suitable event description (in terms of groups) has been selected, membership attribution does indeed become a cognitively straightforward operation. Furthermore, this limitation is by no means unique to Pietraszewski's account. Several authors made the empirical case for each of these interpretive strategies in infancy and toddlerhood (trait-based, Hamlin, 2013; relational, Thomsen & Carey, 2013; group-based, Powell & Spelke, 2013; Rhodes, Hetherington, Brink, & Wellman, 2015), yet explicit discussions about how young learners may adjudicate among these have been lacking.

In fact, Pietraszewski appears aware of this issue. In discussing the "defense" scenario (in which agent C, having witnessed A attacking B, then retaliates against A), Pietraszewski concedes that observers may interpret A's behavior not as indicative of their common group membership, but rather of the agent's sensitivity to moral demands (e.g., punishing unprovoked aggressors, Geraci, 2021; Kanakogi et al., 2017). To determine which of these interpretations is being upheld, Pietraszewski proposes a way to assess the cross-situational consistency of role assignment: If A and C belong to the same group, they should also fill the appropriate group-constitutive roles in the remaining three scenarios; no such consistency should be expected if A's intervention was instead prompted by moral demands. Although falling short of solving the arbitration problem, this verification strategy suggests that different frames may yield representations of different granularity: A trait-based frame, for instance, may not encourage the encoding of the patients' identity (because not essential to predicting behavioral consistency) as much as a relational frame, or, conversely, it may be compatible with a wider range of pairwise combinations than a relational structure based on assumptions of linear ordering (Mascaro & Csibra, 2014).

In sum, although Pietraszewski's sterling contribution to social cognition may not shed new light on the still-neglected issue of how observers can decide among competing frames, it advances a computationally tractable and straightforwardly testable theory of how humans reason about social groups, which developmental scientists, in particular, will undoubtedly benefit from.

Financial support. This research was supported by a European Research Council Advanced Investigator grant (no. 742231 "PARTNERS").

Conflict of interest. The author declares no conflicts of interest.

References

Boseovski, J. J., & Lee, K. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology, 42*(3), 500.
Geraci, A. (2021). Toddlers' expectations of corporal third-party punishments against the non-defender puppet. *Journal of Experimental Child Psychology, 210*, 105199.
Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science, 22*(3), 186–193.
Kanakogi, Y., Inoue, Y., Matsuda, G., Butler, D., Hiraki, K., & Myowa-Yamakoshi, M. (2017). Preverbal infants affirm third-party interventions that protect victims from aggressors. *Nature Human Behaviour, 1*(2), 1–7.
Kazem, A. J., & Aureli, F. (2005). *Redirection of aggression: Multiparty signalling within a network*. Animal communication networks, 191–218.
Mascaro, O., & Csibra, G. (2014). Human infants' learning of social structures the case of dominance hierarchy. *Psychological Science, 25*(1), 250–255.
Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences, 110*(41), E3965–E3972.
Rhodes, M., Hetherington, C., Brink, K., & Wellman, H. M. (2015). Infants' use of social partnerships to predict behavior. *Developmental Science, 18*(6), 909–916.
Thomsen, L., & Carey, S. (2013). Core cognition of relational models. In M. R. Banaji, & S. A. Gelman (Eds.), *Navigating the social world: What infants, children, and other species teach us* (pp. 16–22). Oxford University Press.

How do we know who may replace each other in triadic conflict roles?

Lotte Thomsen[a,b,c]

[a]Department of Psychology, University of Oslo, Oslo, Norway; [b]Department of Political Science, Aarhus University, Aarhus, Denmark and [c]Center for the Experimental-Philosophical Study of Discrimination, Aarhus University, Aarhus, Denmark
lotte.thomsen@psykologi.uio.no | doi:10.1017/S0140525X21001473, e124

Abstract
Group representations need not reduce to triadic conflict roles, although we infer group membership from them. A conceptual primitive of <group> as one solidary, bounded unity or clique

may motivate and facilitate reasoning about cooperative group interactions in context with and without intergroup conflict and may also be necessary for representing which agents would replace one another in a triadic conflict.

The learnability of the social world requires that some core, abstract primitives guide attention to its underlying relational structure so that likely and appropriate forms of interaction (costs and benefits) may be predicted and motivated throughout social life (Fiske, 1991, 1992; Fiske, Thomsen, & Thein, 2009; Sheehy-Skeffington & Thomsen, 2020; Thomsen, 2020; Thomsen & Carey, 2013; Thomsen, Frankenhuis, Ingold-Smith, & Carey, 2011).

Pietraszewski points out that much social psychological, intergroup theory takes for granted the concept of <group> and argues (a) why the concept of group is evolvable and (b) that engagement in primitive triadic conflict roles carries critical information about who is in a group alliance with whom. Indeed, one might argue that the meaning of <group alliance> – or as the author puts it "group in the context of conflict" – *is* <a set of agents who help each other in conflict>. Pietraszewski's careful elaboration of how the meaning of group alliance relates to triadic conflict primitives is thought-provoking and valuable for social and cognitive psychological theory. But it does not yet solve the question of what is the structural representation of the conceptual primitive <group>: Although we intuitively infer the existence of groups based on how people help or hinder each other in conflict, this does not entail that group reduces to engagement in triadic conflict roles and so does not solve what the <group> concept we infer is.

Although between-group conflict likely characterized the evolutionary context of humans (cf. e.g., Richerson & Boyd, 2008), groups coordinate and cooperate in many more ways than antagonistic conflict, and so any universal concept of "any and all groups" should also undergird these forms of non-conflict related solidarity – for instance, cooperative child care or the redistribution of resources within the group to help the young, sick, and old. Hence, what we infer from evidence of triadic conflict roles must be a group axiom which should also undergird the above cooperative practices, rather than a summary concept of <group-in-conflict> based only on triadic conflict primitives. If not, a strict conflict primitives proposal must either imply that "group" basically means something different in, say, the statements of "a group should work together," "a group has to help those who are in need through no fault of their own," "a group has to first take care of its own," and "a group has to stand together against its aggressors"; or, alternatively, demonstrate that non-conflict-based forms of group cooperation are also undergirded by triadic conflict primitives (as, for now, unsuccessfully attempted by the author for the case of reciprocity).

I instead posit that we use the word group in each of the above sentences simply because we mean the same by it – conflict or not – namely, a bounded, communal, solidary, merged, or fused unity of people whom and whose interests we treat as one for the purposes at hand (cf. Fiske, 1991; Fiske et al., 2009; Thomsen & Carey, 2013; Thomsen & Fiske, 2018). Pietraszewski writes that many existing theories about social groups are "simply metaphorical (typically a subsumption or containment metaphor)" (in the abstract), and he argues that such a container metaphor cannot account for shifting or superordinate identities and that the metaphor cannot be literally correct because individual people know that they are not the same individuals. According to Pietraszewski, this demonstrates "the danger of intuitive theoretical reasoning."

Yet a simple straightforward theoretical alternative is that the reason researchers and lay-people alike intuitively and effortlessly use conceptual container, subsumption, and unity metaphors – which are found across language families and cultural practices – to speak about social groups is that these are, in fact, the forms of thought that humans use to reason about groups (cf. Fiske, 1991, 1992; Kovecses, 2010). Of course, cognitive heuristics or summary representations do not have to be literally true to function. And container representations can, in fact, describe the example of shifting social identities where two groups of former enemies unite (sic!) against a third one, by simply subsuming two social containers within a superordinate, common one. This also highlights the convenient computational fact that container schemas embed recursively.

I agree with Pietraszewski that the very act of categorization cannot itself be what group representations are about. Instead, the core, critical information must be the relational features for how costs and benefits are shared in solidary ways within the categorized groups. One theory already on the market which makes this explicit assumption about communal groups – in a manner not bound to conflict roles – is relational models theory and its prediction that universal evolved core relational primitives manifest already in early childhood (Fiske, 1991, 1992; Thomsen & Carey, 2013; Thomsen & Fiske, 2018). It makes the argument that uniting "as one" may foster extraordinary acts of solidary sacrifice, and that we infer and feel the existence of solidary unity based on such sacrifice. Identity fusion theory makes similar points (e.g., Whitehouse, 2018).

Pietraszewski points out that summary group representations must scale up to social identities and assumes that this is possible by simply treating people with the same social identity as interchangeable with respect to the triadic conflict-role slots. However, the key question remains how that is possible: How do we know that people are interchangeable with respect to triadic conflict roles – what is the structure of the cognitive architecture that makes it possible to think of several individual people in this way as "one and the same"? I posit that these inferences are precisely what the concept of group licenses, and that it takes the structural form of a *bounded unity/in-out categorical container* between individuals so that they are treated, and treat each, other "as one," greater than each individual, for the purposes at hand (cf. Fiske, 1991; Thomsen & Carey, 2013; Thomsen & Fiske, 2018). As the peasants in the village Fuente respond in the play *Fuenteovejuna* when they are all tortured to make them reveal which of them killed a military commander: *Fuente did it* (Lope de Vega, 1619/1977).

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing.* Free Press.
Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review, 99*(4), 689.
Fiske, A. P., Thomsen, L., & Thein, S. M. (2009). Differently embodying different relationships. *European Journal of Social Psychology, 39*(7), 1294–1297.
Kovecses, Z. (2010). *Metaphor: A practical introduction.* Oxford University Press.
Lope de Vega, F. (1619/1977). Fuenteovejuna. Editorial Ramon Soprena, Barcelona.

Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution.* University of Chicago Press.

Sheehy-Skeffington, J., & Thomsen, L. (2020). Egalitarianism: Psychological and socioecological foundations. *Current Opinion in Psychology, 32*, 146–152.

Thomsen, L. (2020). The developmental origins of social hierarchy: How infants and young children mentally represent and respond to power and status. *Current Opinion in Psychology, 33,* 201–208.

Thomsen, L., & Carey, S. (2013). Core cognition of relational models. In M. R. Banaji & S. A. Gelman (Eds.), Navigating *the social world: What infants, children and other species can teach us* (pp. 17–22). Oxford University Press.

Thomsen, L., & Fiske, A. P. (2018). Communal sharing/identity fusion does not require reflection on episodic memory of shared experience or trauma–and usually generates kindness. *Behavioral and Brain Sciences, 41,* e219. https://doi.org/10.1017/S0140525X18001784

Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science, 331*(6016), 477–480.

Whitehouse, H. (2018). Dying for the group: Towards a general theory of extreme self-sacrifice. *Behavioral and Brain Sciences, 41*, e192. https://doi.org/10.1017/S0140525X18000249

Group? What group? A computational model of the group needs a psychology of “us” (not “them”)

Janet Wiles[a], S. Alexander Haslam[b], Niklas K. Steffens[b] and Jolanda Jetten[b]

[a]School of ITEE, The University of Queensland, St Lucia, QLD 4072, Australia and [b]School of Psychology, The University of Queensland, St Lucia, QLD 4072, Australia
j.wiles@uq.edu.au | a.haslam@uq.edu.au | n.steffens@uq.edu.au | j.jetten@psy.uq.edu.au | https://researchers.uq.edu.au/researcher/13 | https://researchers.uq.edu.au/researcher/1946 | https://researchers.uq.edu.au/researcher/2864 | https://researchers.uq.edu.au/researcher/900

| doi:10.1017/S0140525X21001266, e125

Abstract

Groups are only real, and only serve as a basis for collective action, when their members perceive them to be real. For a computational model to have analytic fidelity and predictive validity it, therefore, needs to engage with the psychological reality of groups, their internal structure, and their structuring by (and of) the social context in which they function.

Tintin: How did it go JP?

JP: All right.

Gilou: What did you tell Paquin?

JP: Nothing more. I just stood by my statement.

Gilou: You fucker.

Tintin: Hold on JP… You didn't say you'd make a mistake?

JP: You know we all went in after the shots. I'm not lying. I've no reason to stick my neck out for you.

Gilou: If you say that it's homicide and we all get fired.

JP: Not me. You.

[Gilou goes to hit JP, but Tintin intervenes]

Tintin: Stop… stop. For fuck's sake JP, we have to stick together as a team.

JP: What team? What sticking together? You know nothing about me. So don't talk to me about teams!

(Spiral, Series 4 Episode 2).

The above exchange from the French crime series *Spiral* (*Engrenages*) speaks to the fact that from the perspective of “users” the definition of groups is far from straightforward. For Gilou and Tintin – long-time members of their small but maverick police unit – the team is very real. It binds them together and is a basis for them to not just to cooperate, but also to make significant sacrifices on each other's behalf. Indeed, the argument we see here relates to the fact that Gilou and Tintin have just lied to cover up the misdemeanours of another team-member (Laure). But this is something that JP – a newcomer to the team – was not prepared to do. And as he intimates, the fundamental reason for this is that the team is not real for him.

Trying to construct a model (computational or otherwise) of what is going on here is an important question – for all the reasons that Pietraszewski sets out in his carefully argued and well-researched paper. Indeed, for social and computational scientists alike, this is an immensely important challenge. Yet, although there is much to commend in Pietraszewski's attempt to do this, his proposed solution suffers from four core problems.

The first, and most basic problem with Pietraszewski's model is that treats the group as objectively definable based on a set of external exigencies (e.g., relating to conflict). Applied to Gilou and Tintin's team this would suggest that we could develop a computational theory which implied that so long as the team satisfied a set of predefined criteria it would be *a team*. Yet we see from the exchange in the Paris police station that no specification can do this without taking stock of the *psychological reality of group members themselves* – a reality that for Gilou and JP is structured by factors such as status (e.g., as a newcomer or old-timer), accessibility (e.g., prior experience of the group as a platform for collective action), and fit (e.g., a context of threat which makes it a meaningful entity) (Oakes, Haslam, & Turner, 1994).

Following from these observations, the second problem with Pietraszewski's model is that it fails to take stock of the psychological foundations of group life. Before we can model intergroup behaviour, we need first to model the processes that make group behaviour possible (Turner, 1982). After all, as the scene from *Spiral* shows, we can only engage in conflict with “them” if we have first developed a sense of “us” that makes cooperation among ourselves possible (Coser, 1956). Accordingly, models that seek to capture the capabilities and affordances of groups (e.g., as a basis for trust, communication, mutual influence, and organisation; Haslam, 2004) need to start by understanding the foundations of the internalised sense of group membership that gives rise to a sense of *social identity* (Turner, Oakes, Haslam, & McGarty, 1994).

Third, a satisfactory model requires an appreciation of the internal structure of groups and of the norms that regulate their behaviour. Among other things, this is because norms determine group members' understanding of acceptable behaviour and dictate the different ways in which they should relate to each other (e.g., as outlined by Fiske, 1992). For example, what may look like conflict to observers may be perceived as cooperation by its members (and vice versa) and such considerations are essential for understanding whether particular behaviour presages group formation or else disintegration. As Gilou and Tintin note, although a commitment to honesty and openness will make many groups stronger, it may destroy others. Again, this is something we need to appreciate not as onlookers who stand outside the group, but as *insiders* for whom the group is psychologically real in ways that make it a platform for both sense-making and social action.

Developing and integrating these various observations, the fourth problem with Pietraszewski's model is that it fails to recognise that group definition and associated group processes necessarily depend on *social context*. Psychological group membership is fluid not fixed, labile not ossified. This means that "who we are" and what it means to be member of a given group varies across situations – for example, as a function of the groups against which we compare ourselves, the specific tasks and challenges we confront, and the way we are led. A computational model of social groups, therefore, needs to include an appreciation of social context and of its capacity to structure psychological group membership. Without this, the model's predictive power – and hence its efficacy – will be fatally compromised.

Notwithstanding these problems, we see the challenge that Pietraszewski sets out as a worthy one. And certainly the fact that, psychologically, the group is a subjective context-dependent entity rather than a static objective object, does not mean that it cannot – and should not – be modelled computationally. Indeed, the four problems we have highlighted constitute what we see as primitives for an alternative framework that might do precisely this. Critically, though, they suggest that the key challenge for those who would develop computational models of the group is to embrace an understanding of groups that is faithful to the contextuality of group members' subjective experience. In short, we need a model of "us" not "them" that speaks to the richness of social identity (i.e., a sense of "us-ness") as a platform for the potentialities of group life.

Financial support. This work was supported by the Australian Research Council (JW, grant number CE140100041).

Conflict of interest. None.

References

Coser, L. A. (1956). *The functions of social conflict*. Routledge.
Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review, 99*, 689.
Haslam, S. A. (2004). *Psychology in organizations: The social identity approach.* Sage.
Oakes, P. J., Haslam, S. A., & Turner, J. C. (1994). *Stereotyping and social reality.* Blackwell.
Turner, J. C. (1982). Towards a cognitive redefinition of the social group. In H. Tajfel (Ed.), *Social identity and intergroup relations* (pp. 15–40). Cambridge University Press.
Turner, J. C., Oakes, P. J., Haslam, S. A., & McGarty, C. (1994). Self and collective: Cognition and social context. *Personality and Social Psychology Bulletin, 20*, 454– 463. doi: 10.1177/0146167294205002.

A neuroscientific perspective on the computational theory of social groups

Marco K. Wittmann[a,f,g] , Nadira S. Faber[b,c] and Claus Lamm[d,e]

[a]Department of Experimental Psychology, Wellcome Centre for Integrative Neuroimaging (WIN), University of Oxford, Oxford OX1 3SR, UK; [b]Department of Psychology, University of Exeter, Exeter EX4 4QG, UK; [c]Uehiro Centre for Practical Ethics, University of Oxford, Oxford OX1 IPT, UK; [d]Faculty of Psychology, University of Vienna, Vienna 1010, Austria; [e]Vienna Cognitive Science Hub, University of Vienna, Vienna 1010, Austria; [f]Department of Experimental Psychology, University College London, London WC1H 0AP, UK and [g]Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, UK
marco.k.wittmann@gmail.com | nadira.faber@gmail.com | claus.lamm@univie.ac.at | https://sites.google.com/view/marcokwittmann | http://nadirafaber.com/ | https://scan.psy.univie.ac.at | doi:10.1017/S0140525X2100128X, e126

Abstract

We welcome a computational theory on social groups, yet we argue it would benefit from a broader scope. A neuroscientific perspective offers the possibility to disentangle which computations employed in a group context are genuinely social in nature. Concurrently, we emphasize that a unifying theory of social groups needs to additionally consider higher-level processes like motivations and emotions.

Social groups are studied in a variety of fields in the behavioural and brain sciences, mirroring their central role in human and nonhuman societies. As three researchers studying groups and their individuals from different perspectives (e.g., Faber, Häusser, & Kerr, 2017; Lockwood et al., 2018; Rütgen et al., 2015), we very much welcome an attempt for a unifying framework on groups. A shared conceptualization of groups would indeed allow researchers to bridge gaps between disciplines and could be extended to core group topics beyond conflict, such as cooperation and coordination problems. However, to really be unifying, we argue a broader scope than the one presented in the target article would be needed. Such a broader scope should, in particular, consider (1) a neuroscientific angle that (2) incorporates motivations and emotions.

The target article dissects the cognitive mechanisms allowing individuals to navigate social groups. It argues that group computations arise from considering *costs* that one agent imposes on another. Representing costs in group settings presses basic *relational primitives* into action. Relational primitives are the computational scaffold that enables us to think of the relationships between individuals in a group. However, we know that the computational architecture proposed, should it exist, must arise from neural processes. In neuroscience, we routinely consider learning from rewards and losses using computational approaches (Sutton & Barto, 2018). Pietraszewski's framework suggests that this quantitative framework of learning and decision-making might, in fact, be an adequate route to understanding the relational primitives at the heart of social group membership computations.

Neural networks in prefrontal cortex and interconnected subcortical regions have been identified that determine how people weigh costs and benefits in non-social settings (Basten, Biele, Heekeren, & Fiebach, 2010; Klein-Flügge, Kennerley, Friston, & Bestmann, 2016). One intriguing implication from Pietraszewski's framework is that the same machinery that has evolutionary arisen to guide cost–benefit decisions in non-social domains may be used to instantiate relational primitives in group contexts. Should this be the case, then this raises a fundamental question about the social nature of group representations. This question is: To what degree are the component processes underlying our ability to navigate social groups *specifically social*? Neuroscientifically, it is likely that many component processes underlying social cognition are shared between social and non-social domains (Wittmann, Lockwood, & Rushworth, 2018). For instance, the amygdala and the lateral orbitofrontal cortex are important for learning from rewards and associating them with specific stimuli (Murray & Rudebeck, 2018). These computations might underlie some of their contributions to social cognition (Munuera, Rigotti, & Salzman, 2018; Sliwa & Freiwald, 2017). However, other brain regions have been more specifically linked to our ability to think about other people and infer their beliefs such as the temporoparietal junction and dorsomedial prefrontal cortex (Lamm, Bukowski, & Silani, 2016; Saxe, 2006) and it is possible that at least some of the computations performed in these regions are specifically needed for navigating social environments.

Applying this perspective to group computations, we might speculate that group cognition relies similarly on a mixture of social and non-social mechanisms. Following the rationale of the target article, it might draw particularly on the ability to compute rewards and costs that ensue from the actions of ourselves and others. In addition, an ability that seems particularly pertinent for group cognition may be the ability to infer relationships between multiple agents. Recent studies have explored the specific computations via which the brain computes relationships between objects and even abstract concepts (Behrens et al., 2018). This might be central for instantiating the relational primitives underpinning our ability to think about social groups. Nevertheless, despite the potential existence of such a domain general mechanism for forming relationships, it is possible that, in particular, dorsomedial prefrontal cortex, one of the most prominent regions in social cognition research, is specifically important for forming relationships between social agents (Izuma & Adolphs, 2013). Dorsomedial prefrontal cortex represents self and others in an interdependent frame of reference and appears to be causally important to separate out self and other related information (Wittmann et al., 2016, 2021).

Therefore, by employing a neurocomputational perspective, we may gain more precise information on which aspects of group representations may be genuinely social. However, by no means do we suggest being “reductionist” in a computational theory of groups. In fact, there are specific component mechanisms involved in social processes that may not fit in the categories proposed in the target article. Specifically, *motivations* and *emotions* are crucial for different aspects of group functioning – generally, and when it comes to conflict within and between groups. For example, a key social motivation for group functioning is the desire to build a positive reputation in the eyes of other people (cf. Faber, Savulescu, & Van Lange, 2016). This motivation critically shapes prosocial behaviour (Ariely, Bracha, & Meier, 2009; Nowak & Sigmund, 2005) and also group-decision making (De Dreu, Nijstad, & van Knippenberg, 2008; Faulmüller, Mojzisch, Kerschreiter, & Schulz-Hardt, 2012). Regarding emotions, empathy is an exemplary social emotion that is crucial. Empathy allows us to understand each other from a first-person experiential perspective (e.g., Lamm, Rütgen, & Wagner, 2019). Although this can have beneficial effects on prosocial behaviours, such as intergroup cooperation, there is also a potential “dark side” to empathy. This may come out in competitive contexts (when we use our understanding of others to better compete against them), as well as when considering that empathy and the ensuing behaviour is prone to ingroup biases (Bloom, 2017). Although we have only started to understand the neurocomputational processes that underpin motivations and emotions at an individual level, even less is known on how a group context may alter or amplify these processes – or vice versa, how these processes determine membership in groups.

In summary, neuroscience provides a complementary approach that may enrich the proposed computational theory of social groups. It may help determine the precise – social and non-social – component mechanisms underlying group computations and provide a scaffold to incorporate additional component processes relating to motivations and emotions as well as their interaction into a computational theory of social groups.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. The authors declare no competing interests.

References

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review, 99*(1), 544–555, https://doi.org/10.1257/aer.99.1.544.

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences of the United States of America, 107*(50), 21767–21772, https://doi.org/10.1073/pnas.0908104107.

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron, 100*(2), 490–509, https://doi.org/10.1016/j.neuron.2018.10.002.

Bloom, P. (2017). Empathy and Its discontents. *Trends in Cognitive Sciences, 21*(1), 24–31.

De Dreu, C. K. W., Nijstad, B. A., & van Knippenberg, D. (2008). Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review, 12*(1), 22–49, https://doi.org/10.1177/1088868307304092.

Faber, N. S., Häusser, J. A., & Kerr, N. L. (2017). Sleep deprivation impairs and caffeine enhances my performance, but not always our performance: How acting in a group can change the effects of impairments and enhancements. *Personality and Social Psychology Review, 21*(1), 3–28, https://doi.org/10.1177/1088868315609487.

Faber, N. S., Savulescu, J., & Van Lange, P. A. (2016). Reputational concerns as a general determinant of group functioning. *Behavioral and Brain Sciences, 39,* e148. https://doi.org/10.1017/S0140525X15001363.

Faulmüller, N., Mojzisch, A., Kerschreiter, R., & Schulz-Hardt, S. (2012). Do you want to convince me or to be understood? Preference-consistent information sharing and its motivational determinants. *Personality and Social Psychology Bulletin, 38*(12), 1684– 1696, https://doi.org/10.1177/0146167212458707.

Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron, 78*(3), 563–573, https://doi.org/10.1016/j.neuron.2013.03.023.

Klein-Flügge, M. C., Kennerley, S. W., Friston, K., & Bestmann, S. (2016). Neural signatures of value comparison in human cingulate cortex during decisions requiring an effort-reward trade-off. *Journal of Neuroscience, 36*(39), 10002–10015, https://doi.org/10.1523/JNEUROSCI.0292-16.2016.

Lamm, C., Bukowski, H., & Silani, G. (2016). From shared to distinct self–other representations in empathy: Evidence from neurotypical function and socio-cognitive disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1686), 20150083, https://doi.org/10.1098/rstb.2015.0083.

Lamm, C., Rütgen, M., & Wagner, I. C. (2019). Imaging empathy and prosocial emotions. *Neuroscience Letters, 693,* 49–53, https://doi.org/10.1016/j.neulet.2017.06.054.

Lockwood, P. L., Wittmann, M. K., Apps, M. A. J., Klein-Flügge, M. C., Crockett, M. J., Humphreys, G. W., & Rushworth, M. F. S. (2018). Neural mechanisms for learning self and other ownership. *Nature Communications, 9*(1), 4747, https://doi.org/10.1038/s41467-018-07231-9.

Munuera, J., Rigotti, M., & Salzman, C. D. (2018). Shared neural coding for social hierarchy and reward value in primate amygdala. *Nature Neuroscience, 21,* 415–423. https://doi.org/10.1038/s41593-018-0082-8.

Murray, E. A., & Rudebeck, P. H. (2018). Specializations for reward-guided decision-making in the primate ventral prefrontal cortex. *Nature Reviews Neuroscience, 19*(7), 404–417, https://doi.org/10.1038/s41583-018-0013-4.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291–1298, https://doi.org/10.1038/nature04131.

Rütgen, M., Seidel, E.-M., Silani, G., Riečanský, I., Hummer, A., Windischberger, C., … Lamm, C. (2015). Placebo analgesia and its opioidergic regulation suggest that empathy for pain is grounded in self pain. *Proceedings of the National Academy of Sciences, 112*(41), E5638–E5646, https://doi.org/10.1073/pnas.1511269112.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology, 16*(2), 235–239, https://doi.org/10.1016/j.conb.2006.03.001.

Sliwa, J., & Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. *Science (New York, N.Y.), 356*(6339), 745–749, https://doi.org/10.1126/science.aam6383.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT Press. Wittmann, M. K., Kolling, N., Faber, N. S., Scholl, J., Nelissen, N., & Rushworth, M. F. (2016). Self-other mergence in the frontal cortex during cooperation and competition. *Neuron, 91*(2), 482–493, https://doi.org/10.1016/j.neuron.2016.06.022.

Wittmann, M. K., Lockwood, P. L., & Rushworth, M. F. S. (2018). Neural mechanisms of social cognition in primates. *Annual Review of Neuroscience, 41*, 99–118, https://doi.org/10.1146/annurev-neuro-080317-061450.

Wittmann, M. K., Trudel, N., Trier, H. A., Klein-Flügge, M. C., Sel, A., Verhagen, L., & Rushworth, M. F. S. (2021). Causal manipulation of self-other mergence in the dorsomedial prefrontal cortex. *Neuron, 109*(14), P2353–2361. https://doi.org/10.1016/j.neuron.2021.05.027.

Author's Response

More "us," less "them": An appeal for pluralism – and stand-alone computational theorizing – in our science of social groups

David Pietraszewski

Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
davidpietraszewski@gmail.com | https://www.mpib-berlin.mpg.de/en/staff/david-pietraszewski | doi:10.1017/S0140525X22000024, e127

Abstract
The target article is an appeal to allow explicit computational theorizing into the study of social groups. Some commentators took this proposal and ran with it, some had questions about it, and some were confused or even put off by it. But even the latter did not seem to outright disagree – they thought the proposal was mutually exclusive with some other enterprise, when in fact it is not. Unfortunately, scientists studying social groups have not yet avoided the thread-bare trope of the blind men studying the different parts of the elephant: We see mutual exclusivity when we should see complementarity. I hope we can all take the next steps of examining how the different enterprises and approaches within our area of research might all fit together into a unified whole.

R1. Impressions

The target article presents two arguments: (1) the study of social groups has not yet been explicit about what constitutes a group representation in the human mind and (2) the *roles within triadic interaction types* account plausibly describes a group representation within the bounds of conflict. No one really took issue with the first argument. Instead, most of the action surrounded the second. A few commentaries accused the triadic account of being just as slippery as past accounts, while more had questions about its explanatory scope and adequacy. Another handful ran with the account in exciting and interesting ways.

While I'm anxious to get to argument two, it's remarkable that no one took issue with argument one. We've been studying social groups for 100 years, yet there seems to be general agreement that this work has failed to produce a single non-metaphorical description of what constitutes a group representation in the mind.

So why aren't we inundated with plausible alternative accounts of what constitutes a group representation? Marvin Minsky captures the essence of the problem: "You can't look for something until you have the idea of it" (2011). Currently, psychology is largely intolerant of stand-alone computational theorizing, particularly in the absence of accompanying data, and is pinning its hopes on experimental effects that will – like iron filings tossed into a magnetic field – somehow all congeal around a theory of how the mind works (van Rooij & Baggio, 2021).

But this has never worked: Computational theories rarely, if ever, fall out of the data. Instead, they inform what data to look for in the first place (e.g., Chomsky, 1959, 1980; Gardner, 1985; Minsky, 1961; Weiner, 1948/1961). Indeed, the same lesson occurs across the history of science: Yes, you need data to arbitrate between theories. But the data themselves are not sufficient for guiding what questions to ask in the first place. For that, you need independent theory (Heisenberg, 1983; Kuhn, 1962/1970).

This brings us to the third and most important argument in the paper: Researchers studying groups must start tolerating stand-alone computational theorizing (theorizing about what information-processing problems exist and how they might be solved) – which includes concerning ourselves with how the mind solves the reduction problem of group representation (as the target article does). If we don't, then we will continue to (i) confuse experimental effects for computational theories (as I worry some of the commentators did), and (ii) limit the information-processing problems that we investigate to the narrow set that have obvious links to experimental effects or trivial solutions.

Finally, a number of other commentators picked up on the fact that a good computational theory suggests a large number of *other* information-processing problems that must also be solved. But this was sometimes treated as a problem *with* the theory, as opposed to being the *point* of the theory (see Box R1). For example, **Ratner, Hamilton, and Brewer (Ratner et al.)** chide me for leaving "vague how people reason about groups when they are not privy to observing behaviors," deferring "the hard work to future directions." But this is like criticizing a restaurant for only cooking the food and not eating it for you too. I agree that being specific about what the end-state mental representation is highlights a problem. Namely, how that representation can come to be. And I agree it *is* a hard problem (even when it is based on behaviors). But that's the point of the computational theory: Making the problem specific enough that we can begin to tackle it. This is exactly what others do in their commentaries (e.g**., Leibo, Sasha Vezhnevets, Eckstein, Agapiou, & Duéñez-Guzmán** [**Leibo et al.**]).

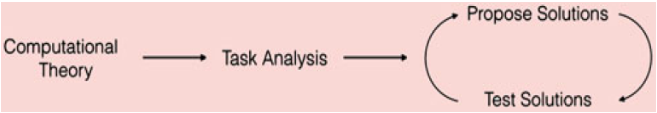
Fundamentally, then, the target article is an appeal for theoretical diversity. Allowing for explicit, stand-alone computational theorizing does not displace other approaches, but compliments them. The title of **Wiles, Haslam, Steffens, and Jetten**'s (**Wiles et al.**) commentary captures it perfectly: "A computational model of the group needs a psychology of 'us' (not 'them')." I couldn't agree more.

Box R1. Problems aren't a problem, they're the point.

Computational theories provide conceptual solutions: they turn abstract, vague notions into information-processing problems. Yet computational theories are not theories of how that problem is solved. Rather, they identify what problem we need, as a community of researchers, to solve. They are the analog of an engineer first analyzing a problem before coming up with a solution to it.

The point of presenting a computational theory is that it allows a larger community of researchers to conduct a *task analysis* of that theory (Marr, 1982; Minsky, 1974). A task analysis involves asking what are the additional information-processing problems that must be solved for this computational theory to be executed in the real-world. That is, the computational theory is broken down into more specific tasks and subtasks. (In the case of the computational theory of scene analysis for vision, for example, subtasks include depth perception, color constancy, object-feature binding, etc.)

Once a task analysis is conducted, researchers start to propose multiple, competing accounts of how these problems may be solved by the mind, and then begin to test for the existence of these solutions. (For example, one problem in color constancy is correcting for reflectance, which in turn requires representing if a surface is smooth or rough, and so on [e.g., Maloney & Brainard, 2010].) This three-step process is depicted below:



R2. Responses to individual commentaries

R2.1. Confusions

Ratner et al. were far and away the most critical. They accuse of me of not holding myself to same standards with which I am criticizing others – which I agree would indeed be deeply unfair, if only it were true. Their argument is that my theory is intuitive because it is not based on empirical evidence: "Pietraszewski provided no empirical evidence to support his assertions. It is unclear why, for instance, he assumes that perceivers inherently view the behaviors in his primitives as evidence for intergroup behavior instead of a string of dyadic interpersonal behaviors. In Figure 3 he circumvents this ambiguity by labeling some positions in the diagram as ingroup and some as outgroup. However, this solution is as tautological as the container metaphor he chastises."

There's a lot to unpack here. First, a false choice: Either participants view interactions as intergroup *or* dyadic behaviors. But the target article argues that intergroup behavior *is* a string of dyadic behaviors (contingent upon a prior dyadic event involving a third agent; a point revisited with **Simandan**). Second, tautology is not equivalent to a lack of computational adequacy, and I only claim the latter about the container metaphor. Third, I agree labeling agents who occupy group-constitutive roles as members of the same group (as I do in Fig. 3) constitutes a tautology – which is why I say so on page 7. But that was the whole point of this section. Figures 3–5 present the theory's definition, and when one presents a definition one is necessarily presenting a tautology. (If it's not clear why, go look up the definition of a tautology.) **Ratner et al.** are confusing the process of defining something with the content of the definition itself, and they present no argument that the definition is itself tautological.

Ratner et al. further criticize me for not providing any empirical evidence. But this criticism misses the point. The target article is a generative theory that points out what evidence we should be looking for in the first place. Ratner et al. also suggest I'm being hypocritical by saying that past definitions of groups are intuitive while my own definition is also intuitive. But they're conflating different senses of intuitive. My concern is not with intuitive theories in the sense of not being based on direct experimental evidence, but with definitional content that requires already having an intuition about the entity being defined. For example, in, "A group exists when two or more people define themselves as members of it," there is little in the definition that does not feed back on to the very notion that is supposed to be defined. It is this sense of intuitive I take issue with, and is the standard that I'm holding myself and others to.

There's more to address in **Ratner et al.** (e.g., they accuse of me sleight of hand when in fact I'm putting my cards on the table) but space is limited, and you get the idea. But I do want to note that later in their commentary they strike a more conciliatory tone: "Pietraszewski's theorizing does not supplant existing work. … We believe that theoretical integration rather than competition between models of 'groupness' is the best path forward."

I agree: The target article is a plea to allow explicit computational theorizing into the study of groups, not at the expense of other approaches, but to complement them. I suspect, then, that we have a case here of misunderstanding born out of interdisciplinarity. Or, as Ratner et al. put it, an integration of a Marr-ian style levels-of-analysis framework with social cognition. These misunderstandings can be seen as the growing pains of that integration.

Two authors, **Levine and Philpot** – without any apparent appreciation of the irony – next attacked me for suggesting that whenever three people get together, the only thing that can happen is a conflict. If that were my theory, it would be a good candidate for the worst theory ever.

To be clear, the account I put forward does not predict that conflict is the only polyadic behavior that can occur. Rather, it articulates what constitutes group membership (to the mind) when polyadic conflict *does* occur. That is, a group representation is in part a representation of roles within polyadic conflict.

The source of misunderstanding may be that the information-processing machinery required to "see" groups is complex, and conflict is but one element of it. A different matter is whether any one group token out in the world is characterized by conflict alone (a distinction that seems to have also tripped up **Ratner et al., Elad-Strenger & Kessler**, and **Thomsen**). As an analogy, we can ask how the visual system represents the world. One element of the visual system are parts for representing lines and edges. That such mechanisms exist is not undermined if we go out into the world and discover that a token of a scene (e.g., an elephant standing in a field) is not exclusively made up of lines and edges.

I do agree with **Levine and Philpot** that conflict protects cooperation. The ethological work out of which the present framework emerged (e.g., Chase, 1985; Strayer & Noel, 1986) makes this very point, and the present account predicts when avoidance of conflict (and therefore conciliation and de-escalation) will in fact occur (see e.g., Pietraszewski, 2016).

Levine and Philpot also worry that I have failed to mention existing work that acknowledges the relational elements of group membership – namely, the meta-contrast principle, which states that: "individuals tend to be categorized as a group…[when]the perceived differences between them are less than the perceived differences between them and other people (outgroups) in the comparative context."

The meta-contrast principle is a lovely description of the categorization process itself, which is, as Bruner, Goodnow, and

Austin (1956), Bruner (1957), and Taylor, Fiske, Etcoff, and Ruderman (1978) put it, the accentuation of between-category differences and the minimization of within-category differences. As such, the meta-contrast principle is a descriptive framework. It acknowledges that categorization is function of context and who is around, but it is not – nor does it try to be – a theory of *which* particular contexts and *which* particular people will be categorized. To get that, you need a theory of what the categorization is *for*.

The target article argues that categorization is relational *because* ultimately what categorization is for is to predict who will take whose side in a conflict. So, I was less concerned with theories that acknowledge that categorization is relational, and more concerned with theories in which the relational property emerges out of the functioning of the system. That said, I agree the meta-contrast principle dovetails nicely with the present account, and is worth including as a way of conceptualizing (and giving a vocabulary for) the context and target specificity of categorization.

R2.2. Questions, including "what about X?"

Simandan finds much to like, but worries that delay, proximity, and loyalty/disloyalty need to be added to the pile of problems to be solved if the present account is to work. I agree. The point of presenting the target article is to provoke exactly what Simandan is doing here: decomposing the account into a number of subproblems. Simandan's deconstruction of the problem of loyalty is beautiful, and should be pushed much further.

Simandan also worries that we do not have a necessary and sufficient theory of groups in the context of conflict. What I meant by *necessary and sufficient* was whether a system that could produce the representations described verbally in the target article would have a representation of groups in the context of conflict. The argument is that it would. A different issue is whether the target article describes everything that you would need to put into a robot to get that robot executing the verbal description – and I would emphatically agree it is not (that's sort of the point). This is what I meant by *computational adequacy*. So I would rephrase this all as that both Simandan and I agree the present account is not computationally adequate, and that an entire universe of the kinds of considerations that Simandan presents needs to be brought to bear if we are to succeed.

Simandan (joining also **Delton** and **Moffet**) also worries about relegating certain group attributes to "ancillary" status. For him, spatial proximity is just too important to be on that list (a point echoed by Moffet).

To clarify, I don't mean intrinsically ancillary, but ancillary with respect to the specific conflict representation in the target article. I'm even happy to stipulate that you might need some kind of proximity representation to get a complete, non-impaired group representation (Lewin field theory comes to mind; see also **Thomsen**). But crucially, space as a concept is not sufficient to produce the kinds of inferences described in the target article (the inference is not that one is literally close; the inference is how agents will interact with one another in terms of costs and benefits).

Delton makes a similar point about cooperation, and I agree. Cooperation is not intrinsically ancillary to groups; it is only ancillary with respect to what counts as a group in the context of conflict.

Delton also worries I'm too harsh on past theories of obligation and interdependence (e.g., Balliet, Tybur, & Van Lange, 2017), and that I've got a few black boxes of my own (such as what counts as a "cost"). While I take the point, I would push back a little on this: The Balliet paper describes evolutionarily recurrent dynamics, but it's not yet a computational theory of something in the mind – and it is only along this dimension that I am evaluating it. I *do* think that it can be turned into a computational theory, and I even take a stab at that in the target article: describing it as a set of modifiers to the polyadic event types, rather than the event types themselves.

With respect to black boxing costs, I emphatically agree, and I'm glad somebody noticed. Theorizing that highlights additional problems to be solved is what we want (a bizarre thing about current psychological theorizing is a desire for descriptions of the mind that don't highlight problems). So, yes, by all means my account – like many others – depends *intrinsically* upon a psychology of representing costs (see also **Wittmann, Faber, & Lamm** [**Wittmann et al.**]), and in no way tackles that problem, aside from highlighting that it *is* a problem. My only other comment, though, is that this is not what the current paper is about: it is about group representations. But to the degree that we have precise theories that do depend on costs (as in this work, and in Delton's other work), we can have more precise theories (and dependent measures) of what constitutes a cost.

Greenburgh and Raihani are similarly constructive in pointing out that cues about what groups exist may be sparse and even hidden, which suggests the existence of additional mechanisms that have to guess about the existence of group-based intentions. Greenburgh & Raihani describe their work on paranoia as a window into how these systems work – including how they calibrate and even break. One notable highlight is that paranoia can produce delusions of the *Alliance* type, which hints at a possible research program in which the triadic interaction types can be used as dependent measures with which to study both normal and clinical-level paranoia.

The tension between the opacity of available cues and the need to predict the future also lies at the heart of **Phillips**'s commentary, in which he is concerned that intentionality may be even more important than my triadic group roles.

While I agree that intentionality is important, I think he has things backwards: The function of the cognitive system is to predict agents' occupation of triadic group roles over phylogenetic and ontogenetic time. Intentions are representations that allow this system to represent if a one-time event is diagnostic of future events. Therefore, it doesn't make sense to say that intentions are more important than the triadic group roles – because what those intentionality representations are pointing to (i.e., the *aboutness* of the intentions) is whether those triadic group roles are likely to occur in the future. Perhaps he thought I was referring to behaviors and not mental representations? (**Phillips** also gets the rock-throwing example wrong: The claim is that all four classes of roles need to be seen or expected; not just one.)

I also think **Phillips** has it backwards when he treats intentionality representations as "thick," and the triadic-roles event grammar as "thin." The systems underwriting the triadic-roles event grammar constitute the bulk of the information-processing. Whereas intentionality representations are icing-on-the-cake representations that either switch some of their calculations on or off. They are placeholders for the fact that appearances can be deceiving.

I liked **Wiles et al.**'s commentary, and not only because of the explicative. They helpfully point out additional problems that need to be solved – although they again seem to think this is a problem *with* my theory, as opposed to a problem *in* my theory (see **Simandan** and **Delton**, above). They highlight four problems, with which I emphatically agree (I also make similar points elsewhere; e.g., Pietraszewski, 2020a, 2020b).

First, members have to experience that some entity is a group ("accessibility") and know in which contexts it is relevant ("fit"). Second, groups don't work if the members don't have an "internalized sense of group membership." Third, "a satisfactory model requires an appreciation of …the norms that…dictate the different ways in which they [group members] should relate to one another." Fourth, "that 'who we are' and what it means to be member of a given group varies across situations."

Again, I agree with all of this. I suspect that **Wiles et al.** think that my robot would be a social imbecile waiting around for some "ossified" set of predetermined "external exigencies," whereas their robot would "get it" and be just fine in the gritty realities of the police locker room. But I think my robot needs what they're describing (to be flexible and to actually engage with specific group tokens), and that their robot needs what I'm describing (to have a representation of the type [group-in-conflict] in the first place).

To be clear, subjective context dependence exists *because* of the problem of having to apply a static, objective inference (the roles within triadic interaction types). An analogy with danger is apt: Suppose that [*X will lead to entropy*] is the mental representation of what it means for X to be dangerous. To be useful, such a representation has to be "shuttled around" computationally (or "protected") from circumstances under which it is not relevant. So if we wanted to build a robot that can see danger, it would need to have additional machinery for being "flexible" and "contextspecific" about when X does and does not lead to entropy (a toaster is perfectly safe, but not when you stick a butter knife in it or bathe with it). The same applies to the *roles within triadic interaction types* account: Different group tokens are going to occupy group-constitutive roles under different circumstances.

For this reason, I'm a bit skeptical about directly pitting objective against subjective, fixed against flexible. After all, objective and fixed information-processing (or developmental) rules cause our flexible, internal subjective states. So our task as scientists is to explain flexibility and subjectivity as outcomes of objective and fixed computational procedures.

R2.3. Clarifying levels of analysis

One way to think about all of this is with respect to the three different levels of reduction or *levels of analysis* at which one can describe the mind (see Pietraszewski & Wertz, 2021): The current framework adopts the middle *functional level of analysis*, in which there are only rule-governed mechanisms. The subjective, flexible experiences that **Wiles et al.** describe then live at the higher *intentional level of analysis*.

Wittmann et al. also make this level of analysis distinction and go on to decompose the problems entailed by the target article into functions (somewhat) known to neuroscientists, noting that (i) "many component processes underlying social cognition are shared between social and non-social domains," and (ii) motivations and emotions are important.

I agree, and have two things to say. First – and I don't think we disagree on this point – motivations and emotions can be understood at all three levels of analysis: We can have a mechanistic account of their computational logic (*functional level*), along with their neuroscientific (*implementational*) and subjective (*intentional*) descriptions. Second, **Wittmann et al.** suggest that "by employing a neurocomputational perspective, we gain more precise information on which aspects of group representations may be genuinely social." I have no doubt that this is true. But I'd offer that such a thing is by-the-wayside. By way of analogy with an automobile, it's helpful to know if turniness is separate from stoppiness. But that's not our final destination in reverseengineering how a car works. So I'd put the enterprise somewhat differently: that we have to specify the processes and procedures that make all of this group stuff work, both in terms of their abstract functional logic, and also their physical realization.

I worry there is an outright level of analysis confusion lurking in **Gelpi, Allidina, Hoyer, and Cunningham** (**Gelpi et al.**); they seem to think that because I'm being concrete I'm both (1) adopting a "bottom-up approach" and (2) claiming that the triadic roles can only be inputs. As such, they think I'm being unnecessarily narrow and leaving out important "top-down" things like categorization, induction, and inference.

But they're wrong on all counts. First, what I'm describing here is the functional role semantics of a group representation, which is orthogonal to the top-down/bottom-up distinction. Second, I *do* think the group-constitutive roles can be inferential outputs in response to abstract category information, which is why I say so in the target article.

The only way I can make sense of **Gelpi et al.**'s comments is if they are confusing intentional and functional levels of analysis. They seem to be adopting a Fodorian view of the mind, in which concrete things are low-level inputs, and abstract things belong to the gooey-center homunculus (see Pietraszewski & Wertz, 2021), whereas in fact concrete things (in the context of a computational theory) are descriptions of everything the mind does at a functional level of analysis – gooey center included.

The funny thing is, I don't disagree with anything that **Gelpi et al.** are saying about categorization or induction. And what they are explaining to me about categorization – such as its being highly flexible – is something I show in my own empirical work. So my issue is with their inferences and argument, and that they seem to be equating categorization with induction. My claim is that the containment metaphor doesn't get you induction. I'm saying, "We don't have a theory of what induction is happening until we known what representations are internal to those induction processes." Gelpi et al. are responding with, "But induction happens!" I know; the problem is how (mechanistically). They seem to use a description of a phenomenon-to-be-explained as an alternative to explaining how the phenomenon happens.

R2.4. Complementary approaches

Fog, Suchow, and **Moffett** likewise say things that are wonderful and that I agree with, but they seem to think is a problem *with* my theory (are we detecting a theme?).

Fog describes research exploring the attributes that make group tokens successful and enduring. This is great, but the target article isn't a theory of that. So I'm confused why he thinks the existence of this research poses a problem for my theory. In fact, the two approaches are complementary: The triadic framework articulates what conflict-related dynamics successful group tokens avoid within their own ranks and provides a way to measure "infighting" (i.e., to what degree group tokens are composed of many other smaller group tokens).

Suchow notes that agents may have diverging representations of what group memberships exist, which requires metarepresentational machinery for representing what others are representing groupwise. He also notes that group representations are intersubjective – that representations need to be somewhat coordinated to get a group off the ground. I agree (similar points appear in the target article). So, if Suchow thinks I wasn't thinking this, I'm glad we could clarify. I also agree that recursive mentalizing is an important and under-studied aspect of social group cognition. I'd add that having more precise accounts of what cues lead to group representations in the first place gives us traction on such mentalizing procedures.

Moffett worries that I'm shoe-horning one particular meaning onto the word "group" – namely, cooperation. (And to think, **Levine & Philpot** were worried that I was only focusing on conflict!)

I agree that "group" is – like any word – a "suitcase" of meanings (as Minsky puts it), and I don't want to lose the suitcase. What about food groups, after all? I also agree that (i) pragmatic and context clues narrow down referents, (ii) that polysemy is a feature and not a bug, and (iii) thinking there's only *one* computational theory for a word is likely misguided. So, I view the enterprise here as the analog of pulling one sock out of the suitcase: but where I'm pulling out a type, not a token.

Moffett – echoing Barth, Sapir, and others – then suggests that large-scale social identities are phylogenetically recurrent entities that also deserve to be called groups, and that socially aligned groups or SAGs (his moniker for the target article's "groups") are somewhat orthogonal to this. I agree and have made a similar point elsewhere (Pietraszewski & Schwartz, 2014).

My only point of disagreement with Moffett is when he suggests that certain group representations "needn't be built on anything more complex…than how we distinguish tiger from panda, with our fear of the other developing toward the former; whether our minds represent any collection of things as a group—humans included—isn't necessarily determined by calculations around whether, and how, they might cooperate."

I worry that we're confusing experience with computation here. If I stand on the ledge of a tall building I experience fear. But that doesn't mean that I'm not calculating that a fall will lead to the entropic disordering of my body. (Or, if you prefer, natural selection "saw" this relationship and tuned the systems that comprise me to intrinsically fear this kind of situation; similar to what the target article called intrinsic ancillary cues.) So it's not clear why **Moffett** wants to jettison calculations about behavior. Does he think that any intact human would fail to understand that members of some feared outgroup would occupy groupconstitutive roles? And what is the fear *about* anyway, if not possible cost infliction?

Oláh and Király similarly point out that there are reasons for the mind to attend to collectives (such as genders, sexes, languages, etc.) aside from inferences about who will take who's side in a conflict. Again, I agree and make similar points elsewhere (Pietraszewski & Schwartz, 2014). They propose calling collectives involving social interactions (of the cooperation/competition sorts; **Moffet**'s SAGs) *social groups*, and collectives that warrant inferences (e.g., sex, language, etc.) *social categories*. Cikara (2021) does a lovely job of articulating this distinction, so I'll simply point to that paper, rather than trying to cram in something here.

What I will say about **Oláh and Király**'s commentary is that it highlights the importance of distinguishing between two different enterprises within our science of groups. One describes the information-processing underwriting each collective studied under the rubric of "social groups." The second describes the information-processing common and universal to the folk-notion "group." The first requires a computationally adequate account of every group token – an array of computational types. The second requires explaining why people think there is an over-arching category "group," and why they agree about a continuum of groupishness across group tokens. (This distinction is well captured by **Ratner et al.**, when they mention Caporeal and Lickel's group types and entitativity continua.)

The target article is concerned with the second enterprise: the representation(s) applicable to the type that captures intuitions about each token being more-or-less-a-group (i.e., the degree of "entitativity"). The argument being that certain collectives are conceivable as "groups" (either by scientists or laypeople) if they have attributes that make them probabilistically informative with respect to conflict expectations (i.e., the event grammar). It's also worth considering whether taxonomies like Lickel's – things like task groups, social categories, intimacy groups, and weak social relationships – describe a factor analysis of tokens, or a set of computational types. It's important work either way, but we shouldn't confuse the two. It would be nice to see work addressing this in the future.

Like **Oláh and Király, Pun and Baron** review relevant developmental work. It's wonderful, I agree with everything they say, and they're doing some of the best current developmental work related to "core" group inferences. So I have nothing more to say.

Likewise, **Cikara** is doing some of the best research on adult's social group representations and inferences, and I'm glad the latent structure learning work is reviewed in her commentary. However, I'm puzzled as to why she's pitting her computational model directly against the target article's computational theory, as they're two very different things.

Cikara's work shows that relative similarity of expressed opinions (how close you and X are, compared to Y) is linked to minimal group or coalitional or alliance-based inferences and motivations (as opposed to absolute similarity on opinions; i.e., how close you and X are). I'm a fan of this work, and it validates past theorizing. But I'm puzzled why Cikara seems to assert that this experimental effect – a computational model, which is a mathematical description of the relationship between dependent and independent measures – is any kind of *computational theory* of a group representation. I also don't understand why she implies that ancillary cues can't be relative. I think they often are, and have been explicit on this point both in the past (e.g., Pietraszewski, 2013, 2020b) and also in the target article.

Finally, in **Cikara**'s effect, *similarity* (of both the relative and absolute kind) is specific to sharing an opinion (what we might call *epistemic coordination*), and is not similarity writ large. I bring this up because distinguishing between different kinds of similarity is crucial, as there are a number of similarity-based theories that simply don't work. For instance, **Ratner et al.** incorrectly argue that I "rely on perceivers inferring similar fate…when analyzing the group-constitutive roles" – when in fact I do not only not do this, I don't even think this works. If you look at the triadic interaction types within the target article, the agents who all share a common fate are those who are attacked. But who's attacked and who's in the same group clearly do not map onto one another. So group membership can't be isomorphic with shared fate.

R2.5. Describing computations

Other commentaries tackle the computations determining which triadic interactions are more or less likely to happen. These include **Fischer, Levin, Rubenstein, Avrashi, Givon, and Oz** (**Fischer et al.**), **Redhead, Minocher, and Deffner** (**Redhead et al.**), **Qi, Vul, Schachner, and Powell** (**Qi et al.**), and **Radkani, Thomas, and Saxe** (**Radkani et al.**).

The **Fischer et al.** commentary asks, in essence, *what are the minimal numerical libraries needed to keep track of and produce social interactions*? They offer a taxonomy covering all possible degrees-of-freedom for social interactions involving costs and benefits – one element of which is Fischer et al.'s specific notion of similarity: subjective Expected Relative Similarity, or SERS, the degree to which I think my behavior will be yoked to yours.

I view **Fischer et al.**'s taxonomy a bit like latitude and longitude coordinates: It describes all possible locations, and the target article describes a particular location. On this account, "group" is a special subset of a larger possibility space. That the target article's proposal fits within this larger possibility space (and that all possible triadic interactions fit within a single figure) is no small feat.

I agree with nearly everything **Fischer et al.** say, but I think they sell themselves short when they say "we do not assume that the applied model embodies a representation of the mind, but expect it to provide testable and valid hypotheses." While they're being careful, I'm happy to speculate (someone has to!) that what Fischer et al. present is represented in the mind – if for no other reason than that it *does* describe a space of all possible social interactions involving costs and benefits.

Redhead et al. add a latent positive tie between the two agents not in conflict within each of the four triadic interaction types. This addition adds tractability from both a network science and on-the-ground-measurement perspective, and opens the door for polyadic benefit-conferral frameworks (something colleagues and I are also starting to work on; e.g., Conroy-Beam, Ghezae, & Pietraszewski, 2021). This notion of ties is helpful in that it allows for dyadic and polyadic continuity, and starts to get at the nature of the underlying mental calculations.

Qi et al. suggest that the triadic framework can be reduced to dyadic welfare trade-off ratios (WTRs) – a representation of how much one agent places another's welfare against their own – and therefore may be a more parsimonious group representation than what is presented in the target article. While I'm sympathetic, I worry this may be acase of what Dennett (1995) called *greedy reductionism*. Yes, you can reduce a computer program to 0's and 1's, but that doesn't mean the program is *just* 0's and 1's. In other words, we shouldn't confuse the ability to describe something in a particular language with the claim that the existence of the language is itself sufficient for the creation of what we just said. (If that were true, we should all stop typing and go home now.)

Here's one argument (of many) for why WTRs probably aren't sufficient: Many species have dyadic relationships. Far fewer have polyadic relationships. Yet both dyadic and polyadic relationships can be described in terms of WTRs. If WTRs are all that are necessary, then we shouldn't see a discrepancy between dyadic and polyadic capacities.

That said, I have no doubt WTRs can describe the triadic interactions (with some exceptions; see Pietraszewski, 2016), if for no other reason than conflict on behalf of another is a benefit to that other (generally), and to impose a cost on another is, well, to impose a cost on another. WTRs are also likely both inputs and outputs to the triadic architecture – that is, WTRs can be calculated as a result of the triadic interactions, and are also values that can inform which agents occupy which roles (I say as much in the target article). But again, we shouldn't confuse the values that the polyadic interaction systems take in and generate as being the same thing as the representation itself.

If instead you want to argue instead that WTRs play out in a particular way in polyadic conflict (i.e., the group-constitutive roles within the triadic interactions), then you've redescribed the present account with different language, which is of course no alternative at all. The upshot of all of this is that WTRs are something different than the target articles' roles within triadic interaction types.

The same comments apply to **Radkani et al.**'s proposal of a recursive utility calculus; that it's a theory of values, not a theory of what gets done with those values, or of how those values coalesce into or map onto a representation of a "group," so it's a category mistake to directly contrast the utility calculus with the proposal in the target article. Otherwise, I agree with pretty much everything else that Radkani et al. say. They note, for example, that groups are a special case of symmetry relationships, but not all relationship will be symmetric – and I agree (indeed, the more general asymmetric cases are covered in Pietraszewski, 2016; a similar observation is made by **Ho, Rosenthal, Fox, Garry, Gopang, Rollins, Soliman, & Swain** [**Ho et al.**]). I also agree that the scaling-up architecture may often get things wrong (a point also made by **Boyer**). I'm not sure they quite understood which elements of the group membership machinery I was claiming could be avoided within the scaling-up architecture, but I don't think either they or I have enough to go on to have that conversation here.

Leibo et al. explore how we might get the representations described in the target article in the head of an organism via reinforcement learning (RL) – but (thankfully) not using RL as an explanation, but as a starting framework for specifying the things you would need to put in the system (or see in the environment) to get the necessary representations (and motivations).

I like the commentary so much that I only have a few things to say. First, I appreciate that they point out that RL does not imply a blank-slate approach. Second, it's heartening to see that the reinforcement learning methods proposed take the credit assignment problem seriously (if you haven't tackled credit assignment, then you don't really have a RL solution). **Leibo et al.** in particular describe a hierarchical structure featuring a manager (who would get rewards from the world), and a worker (for whom the manager creates intrinsic rewards, and who works on more immediate tasks and time-frames). The decoupling between the manager and worker allows for longer time-span credit assignment, among other things. I think this is all great, and also a place for a learnability or task analysis of the environment, where one can ask things like, *What outcomes would the manager be exposed to in principle?* and *What errors or mistakes would it need to protect the learner from?* I also agree that it will be helpful to apply these methods to both real-world and artificial toy world scenarios, and then have the two approaches meet in the middle. I hope the proposed research program comes to fruition.

I also really liked **DeDeo**'s commentary, not least because he suggested my framework may still be too folk psychological (!) in that it may be part of the "talking," "public representation" kind of mental representation – the kind that explains things to others, but that doesn't necessarily do most of the work. DeDeo also mentions sparse coding, a way for system to determine points of maximum inferential leverage at different levels of abstraction. (Essentially, if you and your spouse produce nearly identical social inferences for a third party, then that third party's mind might chunk you and your spouse together for the purposes of making calculations about the social world.) This suggests that my overly granular "atomic-compositional" framework may be better thought of as a population of representations that "mix different levels" of granularity.

I agree with all of this, and have been thinking along similar lines. But to clarify on the first point: The current framework suggests many group summary representations will be private and not communicable (which also means there are far more group representations than first-person phenomenology would suggest). So I might reframe it this way: There are hierarchal and nested sets of representations that chunk the triadic state space in precise ways. This machinery is not hooked up to the "talky" parts of the mind. And the talky parts have a simpler scale-up homogenization architecture, so as to not contaminate the actual computational, precise system.

With respect to sparse coding, I would put it that the atomic system spawns sparse coding representations, which are time-consuming, but once made can be used quickly. So the two are not mutually exclusive. Sparse coding produces efficient "chunked" representations that explain a lot without having to repeat the calculations that made the chunk in the first place. I suspect probable events and behaviors get run through an off-line version of this triadic state-space inference machine (tuned according to current relationships, and modified according to changes to relationships and events). Relationships and events that do not qualitatively shift outcomes are then more readily chunked. In other words, the system is constantly asking "What's the least detailed representation that can be used to generate and predict behavior effectively?"

R2.6. Additional (computational) considerations

Bryant and Bainbridge suggest that social interactions offer readily available cues – either by design, or by accident – for inferring group membership. They note work on auditory communication that suggests yelling and laughter (to take just two examples) have design features that make them surprisingly appropriate for making social inferences of the kind in the target article. This work suggests how the mind comes to wrest group membership representations from an opaque and cue-stingy environment, and is relevant to both the **Leibo et al.** and **DeDeo** commentaries.

Boyer's commentary is profound in two ways. First, he points out that intuitive, folk theories distort our scientific intuitions. Which means that one of the best ways to improve science is to render their operations explicit. I couldn't agree more, and this is one of the themes of the target article.

The second point follows from the first: Boyer asks to what degree the representation [*groups as agents*] is ever veridical, particularly when (1) the scale of interaction goes beyond what the cognitive systems evolved to deal with (e.g., modern nation-states), and (2) when methodological individualism no longer holds (i.e., even if we knew everything about how the mind works, we would still have to capture higher-level social dynamics to understand and predict behavior).

Both points are germane and pressing. I would add that the scaling-up architecture probably also gets things wrong even at relatively small scales (does Vanessa really hate Anne, or is that just a facile inference generated by the fact that Vanessa hates Rachel, and that Anne and Rachel are good friends?)

Furthermore, because everyone shares the same scaling-up architecture, these inaccuracies can become self-fulfilling (Did "America" really want to go to war with North Vietnam? Probably no. Yet it did). The outputs of the scaling-up architecture are also likely easier to coordinate around than the messiness of reality – because the outputs are *for* communication and coordination (see also **DeDeo**). This means that we constantly over-homogenize collectives and their relationships with one another. As much as our group psychology allows us to "see" groups, it blinds us to the many realities underlying them.

Tatone points out what he calls the *arbitration problem*: that there are often competing interpretations (or frames) for understanding any particular event – triadic interaction types included. For instance, *Displacement* may reflect a dominance hierarchy and not a shared group membership. Therefore, computational systems must exist for arbitrating between which event framing is correct (or should be favored).

I agree and would note that what Tatone is expressing here is also mirrored in the commentary of **Radkani et al.**, among others, which is that group relationships are a special class of relationships that determine what kind of triadic interactions are likely. However, they are not the *only* relationships that do so – and we need theories of all the relationships that determine behavior within triadic interactions. Tatone's frames are also the kind of latent representations that will be necessary in any theory of how group representations can come to be and remain accurate (as described by **Leibo et al.** or **Simandan**).

I'd add that I'm not sure that things always get arbitrated cleanly between different possible framings. Given that there are plurality of systems within the mind, there needn't be a single point decision between competing frames – particularly when the contents of the competing frames (e.g., dominance vs. group membership) are not themselves mutually exclusive out in the world. Such simultaneous perceptions may then account for why we can speak, for instance, of ethnic group memberships simultaneously being about group memberships and dominance relationships within a society.

Thomsen suggests that containment, unity, and oneness is a computational theory of groups. I fear this reflects a confusion between types and tokens, and simply retraces what I say about intuition in the target article: "Intuition highlights what is variable about groups (who belongs to what group, and what individual tokens of groups exist) while blinding us to what is universal about group memberships (what constitutes a group, and what is done within cognitive architecture once a group is detected)."

In other words, while I agree with Thomsen on most issues, I think she is pushing us in the wrong direction here: rehashing the intuitive output of an essentialized "oneness" as the computational theory itself – rather than as the intuitive output that needs to be explained by the computational theory. To make just one point: There are an infinite number of ways to treat ourselves and each other as "one." Only a vanishingly small subset generates coherent and functional outcomes. It is our job to specify this subset, not to let intuition silently eliminate all but the coherent and functional.

To be clear, I do agree that group tokens (intuitively) involve containment metaphors and oneness. But these do not describe what is done by the cognitive system(s) once group members are so contained or unified. The target article suggests that containment or oneness corresponds to agents being substitutable within particular event grammars or inference engines. And without this notion of what is done with containment or oneness, there is no computational theory. So **Thomsen** has a perfectly good *intentional level* theory of what constitutes a group (i.e.,

Box R2. Conceptual distinctions and principles raised by the commentaries

Distinction/principle	Applicable commentaries
• Asking for concurrent evidence in a theory of a problem (a computational theory) is no virtue	Cikara; Ratner et al.
• Computational theory ≠ computational model	Cikara
• Behavior ≠ psychological representation	Allen & Richardson; Phillips
• Intentions are not rival with the aboutness of intentions	Phillips
• Problems *in* a computational theory are not problems *for* a computational theory; they're the point	Simandan; Delton; Moffett; Wiles et al.; Suchow
• A claim about an element of a computational type (e.g., how the mind see groups-in-conflict) ≠ a claim about the attributes of tokens (e.g., that groups are only characterized by conflict behaviors)	Allen & Richardson; Greenburgh & Raihani; Levine & Philpot; Ratner et al.; Elad-Strenger & Kessler; Thomsen
• Claims about the mind at the intentional level of analysis ≠ claims about the mind at the functional level of analysis (see Pietraszewski & Wertz, 2021)	Gelpi et al.; Thomsen; Wiles et al.
• Flexibility in the application of a representation is not rival with the fixedness of the representation itself (i.e., that the representation is deterministic and finite and thus scientifically knowable)	Wiles et al.
• Describing a phenomenon is not an alternative to explaining how the phenomenon is produced	Gelpi et al.
• Different and complementary (in approach) ≠ wrong	Fog
• Limits (and thus scope) in a computational theory are a feature, not a bug	Moffett; Oláh & Kir & Király; Ratner et al.; Thomsen
• Theories accounting for group tokens ≠ theories accounting for a continuum of groupishness or entitativity across group tokens	Oláh & Király; Ratner et al.
• Tautology ≠ a lack of computational adequacy	Ratner et al.
• Appealing to similarity writ large is not computationally adequate	Cikara; Ratner et al.
• Being able to describe a representation with a particular framework or language does not mean that the framework or language obviates the representation	Qi et al.; Radkani et al.
• Reinforcement learning is not an explanation of how something is learned; it is a framework for asking what you need to put into a system to get it to learn	Leibo et al.
• The perceived homogeneity of groups is likely amplified compared to their actual homogeneity	Boyer
• Intuition is not a substitute for a mechanistic (functional level) theory	Boyer; Thomsen
• Intuition being the source of an explicit computational theory ≠ allowing intuition to take the place of an explicit computational theory	Ratner et al.

what, at an intuitive level, the "person" represents), but not a *functional level* theory (what, at a functional level, mechanisms represent; Pietraszewski & Wertz, 2021) – which is our concern in the target article.

R2.7. Real-world applications

Finally, both **Allen and Richardson** and **Deminchuk & Mishra** explore how the target articles' *roles within triadic interaction types* account may inform real-world phenomena, such as riot contagion and polarization on social media. They find two elements of the account particularly appealing: that (i) identities emerge out of collective behaviors, and (ii) which identities emerge depends on how bystanders and uninvolved third parties become involved. These commentaries demonstrate how explicit computational theorizing provokes new research questions and ways of looking at old problems.

Allen and Richardson do worry that certain behaviors are not captured by the present account. But this concern amounts to what was already addressed in response to **Levine and Philpot**: that to propose a theory of the conflict-related element of a group representation is not to claim that particular group tokens will be characterized only by conflict behaviors. This, and the distinction between outright behaviors and internal representations, and between ancillary and direct cues (and that the latter can be present in different combinations even with the same group for different members) addresses their concerns – concerns similar to **Simandan**'s in that they (helpfully) point out additional information-problems that have to be solved if the present account is to work.

R3. Conclusion: The study of groups needs more and better conceptual distinctions

In closing, I think everyone who studies social groups understands that we're not only dealing with fascinating science, we're dealing with matters of life and death. Right now, people all over the earth are cowering in fear as they are being bombed and shot at, hurt and abused as a direct result of our species' group psychology. So, as much as we have to get anything right in science, we have to get the psychology of social groups right. In this light, I'm more convinced than ever that what we scientists of social groups need most right now is not more data, nor better methods, but more and better conceptual distinctions. I am grateful to the commentators for starting a collective discussion about what these might be (see Box R2 for a summary). I hope the discussion continues.

Funding. This work was supported by funding from the Max Planck Society.

Conflict of interest. None.

References

Balliet, D., Tybur, J. M., & Van Lange, P. A. M. (2017). Functional interdependence theory: An evolutionary account of social situations. *Personality and Social Psychology Review, 21,* 361–388. https://doi.org/10.1177/1088868316657965.

Bruner, J. S. (1957). On perpetual readiness. *Psychological Review, 64,* 123–152.

Bruner, J. S., Goodnow, J., & Austin, G. A. (1956). *A study of thinking.* Wiley.

Chase, I. D. (1985). The sequential analysis of aggressive acts during hierarchy formation: An application of the "jigsaw puzzle" approach. *Animal Behavior, 1985,* 86–100.

Chomsky, N. (1959). A review of B. F. Skinner's verbal behavior. *Language, 35,* 26–58.

Chomsky, N. (1980). Rules and representations. *The Behavioral and Brain Sciences, 3,* 1–61.

Cikara, M. (2021). Causes and consequences of coalitional cognition. *Advances in Experimental Social Psychology.*

Conroy-Beam, D., Ghezae, I., & Pietraszewski, D. (2021). *A sufficiency test of the alliance hypothesis of race.* Talk presented at Human Behavior and Evolution Society (Virtual).

Dennett, D. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. Simon & Schuster.

Gardner, H. (1985). *The mind's new science.* Basic Books.

Heisenberg, W. (1983). *Encounters with Einstein: And other essays on people, places, and particles.* Princeton University Press.

Kuhn, T. S. (1962/1970). *The structure of scientific revolutions.* University of Chicago Press.

Maloney, L. T., & Brainard, D. H. (2010). Color and material perception: Achievements and challenges. *Journal of Vision, 10*(9), 19–19.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* Henry Holt and Co.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE, 49,* 8–30.

Minsky, M. (1974). A framework for representing knowledge. *Artificial Intelligence.* Memo No. 306.

Minsky, M. (2011). The society of mind, Fall 2011. MIT OpenCourseWare. retrieved from: https://www.youtube.com/watch?v=-pb3z2w9gDg&list=PLUl4u3cNGP61E-vNcDV0w5xpsIBYNJDkU.

Pietraszewski, D. (2013). What is group psychology? Adaptations for mapping shared intentional stances. In M. Banaji & S. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 253–257). Oxford University Press.

Pietraszewski, D. (2016). How the mind sees coalitional and group conflict: The evolutionary invariances of coalitional conflict dynamics. *Evolution and Human Behavior, 37,* 470–480.

Pietraszewski, D. (2020a). The evolution of leadership: Leadership and followership as a solution to the problem of creating and executing successful coordination and cooperation enterprises. *The Leadership Quarterly, 31,* 101299.

Pietraszewski, D. (2020b). Intergroup processes: Principles from an evolutionary perspective. In P. Van Lange, E. T. Higgins, & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 373–391). Guilford.

Pietraszewski, D., & Schwartz, A. (2014). Evidence that accent is a dedicated dimension of social categorization, not a byproduct of coalitional categorization. *Evolution and Human Behavior, 35,* 51–57.

Pietraszewski, D., & Wertz, A. E. (2021). Why evolutionary psychology should abandon modularity. *Perspectives in Psychological Science* doi: 10.1177/174569162199711

Strayer, F. F., & Noel, J. M. (1986). The prosocial and antisocial functions of preschool aggression. In C. Zahn-Waxler, E. M. Cummings, & R. Iannotti (Eds.), *Altruism and aggression: Biological and social origins* (pp. 107–131). Cambridge University Press.

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology, 36,* 778–793.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science, 16*(4), 682–697.

Weiner, N. (1948/1961). *Cybernetics: Or control and communication in the animal and the machine* (2nd ed.). MIT Press.