Presented at: The Social Brain. Biology of conflicts and cooperation
World Forum of Cultures: From Conflict to Cooperation.
Proposals from the Cognitive Sciences.
Barcelona. 17-20 July 2004.
http://www.barcelona2004.org/eng/banco_del_conocimiento/dialogos/ficha.cfm?idEvento=164
To appear in: Villaroya, O & Valencia, L. (Eds.)

Spare Me the Complements:
An Immoderate Proposal for Eliminating the "We/They" Category Boundary

Stevan Harnad

Chaire de recherche du Canada
Centre de neuroscience de la cognition (CNC)
Université du Québec à Montréal
Montréal, Québec, Canada H3C 3P8
harnad@uqam.ca
http://www.crsc.uqam.ca/fr/index2.html

and

Department of Electronics and Computer Science
University of Southampton
Highfield, Southampton
SO17 1BJ UNITED KINGDOM
harnad@ecs.soton.ac.uk
http://www.ecs.soton.ac.uk/~harnad/

Certain biological facts are undeniable: Any creature born with a tendency to ignore the calls of nature -- not to eat when hungry, not to mate when horny, not to flee when in harm's way -- would not pass on that unfortunate tendency. Such a creature would instead be the first in a long line of extinct descendents. Maladaptive traits are eliminated from the gene pool by the very definition of what it means to be maladaptive.

An indifference to one's own survival is such a maladaptive trait. There are daredevils and heroes and saints, but even they differ from the mean only by a bit: They still thrash for breath if deprived of air, and trample their fellows to escape encroaching flames.

So in seeking unselfishness and cooperation, don't aim too high. Basic creature-needs need to be met before we can talk about sharing or sacrifice. There are exceptions. We all know about "inclusive fitness" ("I will lay down my life for 2 brothers, 4 cousins," etc.), but that's all in the family, so although the rule may be "every gene for himself," there are causal dependencies between genes, sometimes even when they are in different bodies, and a mother is better off feeding and sometimes even sacrificing herself for her babies, rather than eating them, if she is to pass on that tendency -- or any tendency at all.

Then, besides selfish genes helping out their clones in kin, there is the question of kin detection itself. We are not equipped with genetic kin-detectors. We must learn who are kin, and the main way we learn this is from early exposure. We imprint on those we live with and see very often from birth, favoring them for sharing and disfavoring them for mating. (Sibling competition for goods is another matter, one in which assertive egoism -- kept in check by older kin -- is probably more adaptive in the very earliest years of life than lackadaisical sharing.)

But once you've opened the door to learning, you've already let in non-kin, for "kin" becomes defined by experience rather than by genetics. The most probable primary care-givers are genetic parents, but not necessarily; and the probability shrinks as you move beyond the nuclear family. Locally, there is no great risk, because even those who are not one's kin are still one's kind, with shared experiences, goods, interests, and, most important, shared enemies. It is not at the boundary of the gene, the organism, the kin-line, or the local kind that cooperation ends and conflict begins, but at the frontier with the enemy kind.

Frontiers shift, however, and enmity comes in degrees: In some cases the enemy of my enemy becomes my friend. Common interests create strategic alliances; enemy may even become kin-in-law. But there is always some boundary, always a current "we" versus a "they," an "in" versus an "out." Is there any way to eliminate those boundaries too?

There may be an answer from category-learning theory: Learning who are one's kin and one's kind are just special cases of the learning of categories (kinds) in general. Animal and plant species are kinds; so are many natural objects and artifacts; so are many actions, events and states (running, walking, thunderstorms, droughts, days, nights) including feeling-states (fear, fondness, longing, loathing). All that's needed to pick out a category is a sample of what is in the category and what is not in the category.

A dichotomy is the simplest form of category (e.g., male/female), but most categories have more than two possibilities: There are many kinds of animals, not just zebras versus giraffes. But whether a category is dichotomous or polychotomous, there is always something basically binary -- indeed categorical -- about categorizing itself. For with every category, a particular instance is either in that category or not in it. That is why the so many English nouns and adjectives have a "non-" or "un-" version:"member/nonmember," "athletic/unathletic," "white/nonwhite," etc. The "un-" refers to the complement of the category, and every well-defined category needs to have a complement -- the set of things that are not in the category. For if a "category" has no complement (or you do not know what its complement is), then it is not a category (or not a category that you yet know).

The reason the complement of a category is so important is that it is the invariant features of members of a category that determine what is and is not a member of that category, and the only way for a cognitive system to find out what those features are is to sample both the category and its complement, in order to learn to detect which features will reliably tell them apart. It must detect the invariance in the variation between members and nonmembers.

For some categories we are already born with the invariant-feature-detectors: We probably don't have to learn the difference between a friendly and a threatening face, and we definitely don't have to learn the difference between being hungry and nonhungry, or between pain and pleasure (though learning may modulate the boundaries somewhat in particular cases). Color categories (red, green, blue, etc.) are probably innate as well, although there too there may be some room for some modulation by experience.

But most categories we have to learn from experience: Open up a dictionary and you will find mostly content-words (nouns, adjectives, verbs, adverbs) that are the names of categories, most of them learned rather than innate categories: How did we learn them? By sampling positive and negative instances (i.e., members and non-members of the category) and getting corrective feedback as to whether we have categorized them correctly or incorrectly. A good example would be learning which kinds of mushrooms are edible and inedible. The corrective feedback could come from an instructor, correcting us as we try to sort samples by trial and error, or the feedback could come from our own digestive systems, as we sample a bit of mushroom and become a bit sick from some kinds and not others.

Mushroom-sorting already gives a hint of how it might be adaptive to find a better way to learn to categorize things than by trial-and-error sorting and its natural consequences, as in the case of tasting mushrooms without an instructor: Trial-and-error learning with corrective feedback is not only time-consuming, but risky. It would be much better if an instructor who could already detect the critical features could spare us the tasting and stomach-aches, or, better still, could tell us explicitly, in words, what the distinguishing features are.

For this, we had to evolve natural language, whose most basic function is to allow us to learn new categories from explicit verbal descriptions instead of just from implicit feature-learning guided by feedback from trial-and-error sampling of members and nonmembers.

But members and nonmembers there must be, in any case, otherwise there are no distinguishing features, hence no category. And if we are to learn the category from a description instead of from direct experience, the distinguishing features must not only exist, but be known to the describer, who must be able to put them into words.

So what if a category had only positive instances? Would there be any way to learn it, either from direct experience or from a verbal description? First, what would be an example of such a category? It would have to be one in which the only thing we can sample is what is in the category, not what is not in it. Its complement must be either unreachable for some reason, or empty.

Let's consider the unreachable case first: We had noted that some feeling-categories are innate: We don't need to learn the features of pain vs. pleasure; we are born already able to detect them (and we approach/avoid accordingly). Let us now consider other feeling-categories:What does it feel like to be a bachelor? Does any bachelor who has never experienced being married know? He may guess, filling in the missing complement that he has not actually experienced either by analogy and extrapolation from approximations to the married state that he has already experienced, or by imagination. And he may be right; he may have guessed the critical features that distinguish what it feels like to be a bachelor vs. what if feels like to be married. But let us also admit that he might be wrong, so that if and when he actually does get married, he might say: this not what I thought being married feels like, so it is only now that I really know what being a bachelor feels like.

The married state, however, is not a feeling category that is unreachable in principle, just one that may not be reached by some, in their experiential lifetimes. Perhaps a few words of wisdom from someone who knows what it's like to be married would have done the trick too, conveying the distinguishing features. But to get an idea of what it is like to deal with a truly uncomplemented category, we have to consider other feeling categories:

What does it feel like to be awake? You might feel that you know, and that this is a perfectly well-defined category, but is it? You have sampled various degrees of alertness and drowsiness, but those differences are differences in degrees of awakeness. They don't define the boundary between being awake and being unawake: You have also experienced the onset of being awake. You know what it feels like to be awake when unable to remember feeling anything immediately before. But that is still just another example of what it feels like to be awake. It is self-contradictory to say that you know what it feels like to be unawake, because, by definition, when you are unawake you are not feeling anything at all.

Never mind. We are not too handicapped by the fact that we don't really have the category "what it feels like to be awake," because the relevant distinctions -- the differences that make a difference in our lives, like the difference between an edible and inedible mushroom -- are the differences in degrees of awakeness ("I am too tired to drive," "I need to get some sleep") and not differences based on what it feels like to be awake: The category awake/asleep is then decided on an objective rather than a subjective basis: I don't know what it feels like to be asleep -- I am omitting dreaming, which is in fact a special form of awakeness: we are speaking here of dreamless sleep, when you are gone, and no one is there, feeling anything -- but the awake/asleep distinction can be made in objective, behavioral, 3rd-person terms. Everyone knows the difference between a moving, responsive person and a snoring, unresponsive one. The rest is what it feels like to have oneself only just transited from the unfelt to the felt state.

You may be wondering what all this has to do with unselfishness and cooperation. I ask for just a little more patience. Consider just one further case, this time not a subjective but an objective one: the category of something that exists. Philosophers have always had problems with that category. It is easy to talk about things that you can and cannot see in the street: You can sometimes see zebras, but never unicorns. A zebra is a horse with stripes, and those kinds exist; a unicorn is a horse with a horn, and those kinds don't exist. That's a perfectly well-complemented category. We all know what a horse, stripes and horns look like, and what you do and don't see in the street.

But what about a drawing or animation of a unicorn? or that horse with a horn that I have no trouble imagining, even though I will never see a live one in the street? What do I mean by saying "that" doesn't exist, when I have just done talking about it, and imagining it in my mind (or drawing it on paper). The feature "never appears in the street" is a perfectly adequate feature for distinguishing kinds of things that do and do not appear in the street, but surely "nonexistent" is more than just that: Can I say of anything that I can imagine and talk about and describe that it does not exist? Or does everything I can imagine and talk about exist, so that the rest is just about what other features it has, besides existing. For "existing" itself is a feature shared by the members of all categories, fictional and nonfictional, visible and invisible, but all alike conceivable.

Last exercise, before we return to conflict and cooperation: What if I tell you that that thing over there (a zebra) is a "laylek"? And that too (a drawing of a uncorn). So is running, walking, red, green, pain, pleasure, friend, foe, and what it feels like to be a bachelor, or married, or awake: All of those are members of the category "laylek". Get out the dictionary and pick any content word, any object, property, action, event, relation or state, abstract or concrete -- they are all members of the category "laylek." Anything and everything is a laylek: Do you know what a laylek is? Do you have any idea of what the boundary between a laylek and a non-laylek is? What the distinguishing features of a laylek are?

Wouldn't it be nice if there were some way to make people just as oblivious to the boundary between own kin/kind and other kin/kind, between "we" and "they"? some means of making "us" into an uncomplemented category like laylek, for all of us?

Note that we are speaking here about the 1st-person plural, not the 1st-person singular. What we are contemplating here is not some mystical dissolution of the boundary between "self" and "other," for that would take us down a biologically maladaptive path, which, as noted earlier, leads to a dead end. My pain/pleasure must continue to be mine, felt by me, not blending into some nebulous shared collective consciousness. So we are not talking about "I" but about "we":

Who/what are "we"? Like every category, this changes with context: In a room in which there are women and men, "we" might be the men vs. the women, or, if there are children and adults, age-groups might be the way identifications and allegiances align. Redistribute the same people in a bigger population, say, mostly very aged people, and the adults and children might all coalesce into the single category "we, the young," even though it was age that had divided them in the first context. Ethnic kind, nationality and local neighborhood work this way too, generating ever-changing groupings, depending on which distinctive feature you are sorting on (and it could be anything).

But the original we/they sorting was clearly the local kin-group -- parents, children, grand-parents, aunts, uncles, siblings, cousins -- vs. non-kin. The infant does not need to learn the boundary between itself and the outer-world, but it does need to learn who are "kin" and who are not. I put "kin" in scare-quotes not only because the distinction is not really based on genetic screening, but because it is so circumstantial and context-dependent: The infant imprints on its early care-givers and familiars, and they are not only the ones who are thereafter perceived as kin, but that early interaction, which occurs in a finite critical period in early infancy and childhood, then becomes the model for all later we/they distinctions.

A child deprived of affectionate early care-givers is a child susceptible to all sorts of later social and psychological problems, so let us not even contemplate depriving children of this all-important early human contact. But then how to finesse the we/they distinction?

We must again return to category learning: The way that we learn to categorize is that in our brains there are learning mechanisms -- perhaps certain kinds of neural nets -- that are adept at analyzing sensory input in order to detect <u>invariants</u>. Recall that the simplest category is a dichotomy: The members of the category are the positive instances, the nonmembers (i.e., the members of the category's complement) are the negative instances. What the brain's invariance detectors learn to do -- under the guidance of corrective feedback from the consequences of categorizing correctly or incorrectly (e.g., eating an edible or an inedible mushroom) is to filter out all the variation from instance to instance that is not correlated with being a member or nonmember, and to extract only the features that are correlated with being a positive and not a negative instance.

These are the distinguishing features of the category, and these are the features that it was impossible to extract in the case of feeling awake or knowing what a laylek is, because all instances are positive: There is nothing that distinguishes what it feels like to be awake, or what a laylek is, from what it feels like not be awake, or what a non-laylek is. There are instance to instance differences in what it feels like to be awake, but those are just variations among positive instances, exactly the features that an invariance-detector learns to <u>ignore</u>, as it learns to distinguish the positive instances from the negative. The same is true with all the things that are layleks: Everything is a laylek, just positive instances. So there is no way to extract an invariance from all that variation. It is not that all things may not have features in common: they may have many features in common. But an invariance must not only be present in the members of a category, but it must be absent in nonmembers. And an uncomplemented category has no nonmembers.

So if we are agreed that we cannot deprive an infant of care-givers, how can we deprive it of the we/they distinction? There is a simple way, though it will sound shocking, perhaps even Orwellian or worse: Rearing the child in aggregates-in-flux instead of an invariant nuclear family throughout its critical period. An aggregate-in-flux is a population of care-givers and age-mates (their optimal number to be determined empirically -- one male and one female, plus a few age-mates being the default option) into which an infant "rotates" (for an interval to be determined empirically, perhaps a few days, perhaps a week or two, but not longer) from the time of its birth till the critical period (again to be determined empirically) for imprinting and the we/they distinction is over (perhaps puberty, perhaps later, perhaps earlier).

During that critical period, in normal child-rearing, the child not only imprints upon and forms life-long attachments to particular people (mostly kin), but the distinction between kin and non-kin becomes the model for later We/They distinctions, distinctions that always involve a discrimination in favor or some (We), and against others (They). But in the aggregates-in-flux, what is there for the child to imprint on? The invariants are there: affectionate care-givers, age-mates, but all the other particulars keep varying. If there is anything at all in all this variance for the child to imprint on, it is only that all these changing care-givers and age-mates are all human beings. That is the only invariant. But it is an invariant that all people share, so it is uncomplemented (or the only boundary it marks is human/nonhuman -- which is not quite optimal for vegetarians like myself!)

Apart from the problem about the species boundary, there is some uncertainty about whether and when the "window" for attachment closes: Is the critical period for attachment ever quite over? When can the flux stop and ordinary life begin? One can keep a female rat lactating all her adult life by constantly giving her new pups to suckle, but would there be any way to segue aggregates-in-flux into a world where individuals do become constant and recurrent at some point? If the attachment window is still open then, then the We/They boundary could still erupt then. We go on forming categories all of our lives; but perhaps later attachments are not as fiercely partisan as early ones.

Would rearing children in aggregates-in-flux during their critical period for kin-attachment succeed in eliminating the We/They boundary? We'll never know, because our selfish genes bias us against even trying such an immoderate proposal. Parents want to keep and form attachments to their own children -- naturally enough, we are inclined to say, but here we are contemplating whether nature can be improved upon, in view of some of its less admirable outcomes.

Perhaps some less radical approximations to aggregates-in-flux are possible during the critical period, but it is hard to see what they could be: Orphanages are unhappy places, serial foster homes are probably too small, change too slow, and the contrasts with normal family life too intrusive;, and kibbutzim are merely bigger families. None provide the requisite unrelenting flux; both still allow imprinting on invariant individuals.

Even if the We category is bound to be complemented in natural life, thought experiments like these might give us a little more insight into how and why it is an invariant of social life, and whether Darwin might have found a better, kinder way.

References

Harnad, S. (1987, unpublished) Uncomplemented Categories, or, What is it Like to be a Bachelor? 1987 Presidential Address: Society for Philosophy and Psychology. http://cogprints.soton.ac.uk/documents/disk0/00/00/21/34/index.html

Harnad, S. (2003) Categorical Perception. Encyclopedia of Cognitive Science. Nature Publishing Group. Macmillan. http://www.ecs.soton.ac.uk/~harnad/Temp/catperc.html

Harnad, S. (2003) Symbol-Grounding Problem. Encylopedia of Cognitive Science. Nature Publishing Group. Macmillan. http://www.ecs.soton.ac.uk/~harnad/Temp/symgro.htm

Harnad, S. (2003) Cognition is Categorization. UQaM Summer Institute in Cognitive Sciences on Categorization. http://www.ecs.soton.ac.uk/~harnad/Temp/catconf.html