ELSEVIER

Cognition 63 (1997) 227–233

COGNITION

Discussion

Teleological reasoning in infancy: The infant's naive theory of rational action
A reply to Premack and Premack

György Gergely[a,*], Gergely Csibra[b]

[a]*Sub-Department of Clinical Health Psychology, University College London, Philips House, Gower Street, London WC1E 6BT, UK*
[b]*MRC Cognitive Development Unit, London, UK*

Abstract

We argue that Premack and Premack's criticism of our demonstration (Gergely et al., 1995) of interpreting goal-directed action in one year-olds in terms of the principle of rationality are ill-founded, and their suggested alternative test for goal-attribution is open to lower level interpretations. We show that the alterative model they propose for our data in terms of 'appropriate' change of means action is but a somewhat imprecise restatement of our account of the infant's naive theory of rational action. Finally, we elaborate and clarify our model of the teleological stance in infancy which we suggest is an as yet nonmentalistic precursor of the young child's later emerging causal theory of mind.

In their discussion of Gergely et al. (1995) David and Ann James Premack (P&P) make three major points (Premack and Premack, in press): (1) They propose an interpretation for our study according to which it ''demonstrates that the infant has attributed a goal to the object'' (p. 8) and suggest that ''the authors' data accrue interest because of the contribution they make to the question of whether infants attribute goal-directedness'' (p. 4). (2) They claim that the above interpretation is ''strikingly dissimilar'' (p. 8) from and superior than our original interpretation of the results. (3) They criticize our treatment of rationality as having ''consequences which are both bizarre and unacceptable'' such as ''completely eliminating the category 'irrational' or 'nonrational''' (p. 10). Let us take up these points in turn.

(1) We can only but applaud P&P's claim that our results indicate goal-attribution in one year-olds, in fact, this is exactly what we have concluded: ''The

* Corresponding author.

PII S0010-0277(97)00004-8

results demonstrate that... 12-month-old infants can identify the agent's goal and interpret its actions causally in relation to it'' (Gergely et al., 1995, p. 165).

(2) Obviously, however, there would be no cause for spilling ink if P&P believed that their interpretation in terms of goal-attribution is identical to ours. In fact, they go on to spell out their model and contrast it to ours as they see it. However, this contrast builds heavily on their construction of a straw-man version of our position which is demonstrably a far cry from our published views.

The straw-man P&P contrast their account in terms of simple goal-attribution (uncontaminated by any notion of rationality) with our alleged (less parsimonious and unmotivated) account according to which the data show ''that the infant has attributed rationality, intentionality, want and belief'' (p. 8). This characterization, however, is simply contradictory to the detailed theoretical discussions in our paper (pp. 172–3, pp. 188–9) in which we explicitly claim that our results do *not* imply the attribution of causal mental states such as intentions, desires or beliefs.

Thus, to clarify what we mean when saying that ''the infant takes the intentional stance'' in our experiment, we wrote: ''... what our study implies is not more than the fact that the 1-year-old causally interprets the actions of the agent in relation to a goal.... In other words,... the infant can come to represent the agent's action as intentional without actually attributing a mental representation of the future goal state to the agent's mind... Thus,... the findings seem sufficiently explained by the hypothesis that the infant applies a paradigm of 'teleological causality'... to interpret the action'' (p. 188). Similarly, concerning beliefs we stated: ''... we do not believe that our results necessarily imply that the infant actually attributes to the other's mind... mentally represented knowledge structures.... [Rather]... the constraints embodied in the infant's naive theories of the world can function similarly to later beliefs in providing background assumptions for evaluating the rationality of an action, *without* having to have the status of a truth-functional propositional attitude concept as later beliefs have...'' (pp. 188–9).

Our real position What we proposed instead is that the infant applies an as yet nonmentalistic interpretational stance in which s/he represents action as being teleologically related to some future goal-state. Furthermore, we have argued that attributing a goal to an action is not simply a function of interpreting certain cues (such as equifinality of outcome), but it also crucially involves the evaluation of the rationality of the goal approach in relation to the constraints on action imposed by the context of reality. In other words, we have indeed claimed that goal-attribution is inherently related to the perceived rationality of the means action relative to a set of background conditions.

In contrast, P&P believe that ''rationality is not... a factor that plays any role in an infant's attribution of intentionality... [and that]... it is goal or goal-directedness... [that is]... the primary factor'' (p. 4). Let us, therefore, compare P&P's model with our real position in the light of the data they discuss. When trying to specify P&P's theory of goal-attribution, however, one finds that they actually propose two different versions of their model. We shall argue that Version 1, which is indeed different from our theory, fails to account for the data and runs into serious difficulties on several grounds. To accommodate these P&P then

modify their model into Version 2, which, however, as we shall try to show, is a somewhat imprecise restatement of our model with some terminological variations to avoid the use of the concept of rationality.

Version 1 of P&P's model: Purely cue-based goal-attribution P&P propose a set of basic conditions that ''cause an infant to attribute goal or goal-directedness to the action of an object'' (p. 4). They identify three properties (trajectory, target and greater than default values) claiming that these represent the ''best available answer'' currently to this problem. In this model goal-attribution is indeed independent of the rationality of the means action: a goal is attributed purely on the basis of certain cues that are inherently salient to the infant.

P&P suggest two tests ''to determine whether or not an infant has attributed a goal'' (p. 5) and compare the predictions of their model with ours in relation to these. The first consists of repeatedly presenting a self-propelled black ball moving towards a red ball. In the test phase, the location of the red ball is changed and on some trials the black ball adjusts its trajectory to head towards the red ball again, while on other trials it does not. According to P&P if infants look longer at the latter type of display, the test proves that they ''have attributed a goal to the black ball. That of contacting the red ball'' (p. 6). They refer to unpublished data indicating that 11 month-olds do, in fact, react in this manner (p. 7).

P&P characterize this test as ''an admirably simple one, so simple, so much the 'shortest possible path' to deciding whether or not an infant has made the attribution of goal, that [they] expect every rational experimenter to use it'' (p. 5). True, we have not done so. However, while we certainly appreciate ''the admirable simplicity'' of this test, we see a number of good reasons why a rational reviewer would have to reject the unpublished data in question as supporting evidence for goal-attribution. For example, the same result could be predicted on a purely associationist basis as during habituation the infants may have formed an association between the terminal position of the black ball and the red ball it contacts. The expectation based on this association is violated during the test phase when the black ball does not change its trajectory (but not when it does): hence the looking-time differences.[1]

Fortunately for the rational experimenter, however, P&P also propose ''another test of virtually equal simplicity'' (p. 7) which is indeed free of the above methodological problems. (In fact, this is the very test we have performed in Gergely et al. (1995).) In the habituation phase of test 2, the black ball in its course to contact the red ball repeatedly surmounts a barrier. During the test phase, the barrier is removed and the black ball either ''immediately accommodates the change, moving directly to the red ball'' or it ''continues to carry out actions previously imposed on it by the barrier'' (pp. 9–10). According to P&P, ''if infants

[1] A further problem for interpreting the results as indicating goal-attribution is that a similar conclusion could be reached even if the *opposite* pattern of looking times were found. In that case it could be argued that the infants attributed the final *location* of the black ball's original movement as the goal. The infants would then be surprised to see the black ball changing its course during the test phase as it would end up at a different location than its assumed goal.

are surprised by the 'failure' of the black ball to adjust to the removal of the barrier, and therefore look longer on these trials... [they]... have attributed a goal to the black ball'' (p. 8).

While we entirely agree with this conclusion, we must point out that it cannot be derived from P&P's cue-based model (Version 1) which considers goal-attribution as purely a function of salient cues (such as trajectory and target). Note, however, that these cues have not been altered by the removal of the obstacle during the test phase. The problem is, therefore, that P&P's model (Version 1) can predict neither that the black ball has to ''adjust to the removal of the barrier'' (p. 8), nor what particular type of adjustment is expectable given the attributed goal and the changed situation. In contrast, the principle of rational action allows for both types of prediction. In our model, for goal-attribution it is not sufficient to identify a goal on the basis of cues (such as equifinal outcome); additionally, the particular manner of goal-approach has to be seen as justifiable within the context of reality constraints. Therefore, if there is a relevant change in the context of reality (e.g., the removal of the obstacle) our model predicts a corresponding adjustment based on the rationality assumption which specifies that a goal-directed action will involve the most rational means action currently available.

P&P then discuss our control condition in which no obstacle is present during habituation (see their Figure 1B) and where the difference in looking times to the old jumping approach vs. the new straight-line approach in the test phase disappears (Gergely et al., 1995). P&P concludes (just as we did) that ''the infants in the control group did not attribute a goal to the small ball'' (p. 9). But why not?

At this point P&P wave their hands and suggest that this is an ''empirical question'' as ''there is no theory that permits predicting which trajectories will cause an infant to attribute a goal'' (p. 9). This is somewhat baffling given the fact that they have just proposed such a theory which they portrayed as ''providing the best available answer to this question'' (p. 4). In fact, they applied their theory to predict the attributed goal in their two proposed tests where the goal was identified on the basis of cues such as trajectory and target. Since, however, the cues in question in the experimental (barrier) condition are identical to those in the control (no barrier) condition, P&P's purely cue-based model (Version 1) is forced into predicting the same goal for the control as it did for the experimental group (see Gergely et al., 1995, p. 186). As such, the model cannot explain the differential results of the experimental vs. the control conditions.

In fact, we know of one other published model as well which generates a specific prediction (different from P&P's) about goal-attribution for the control group: namely, our own theory (Gergely et al., 1995). We proposed that goal-attribution is not simply a cue-governed process, but is also a function of evaluating the rationality of the goal-approach in contrast to other potential alternative actions that could lead to the same goal given the constraints of reality. Since in the no-barrier control condition there clearly is a more justifiable alternative (the straight-line approach) leading to the terminal position of the object's movement, it is predicted that location will not be attributed as goal (a prediction borne out by the results).

Version 2 of P&P's model: The assumption of 'appropriate adjustment' of means actions But P&P might argue that we are also creating a straw-man when we equate their position with Version 1 above. After all, they go on to significantly enrich their purely cue-based model when they identify as ''a fundamental entailment of what is meant by 'having a goal'' (p. 6) the assumption that goal-directed agents are capable of ''flexible appropriate adjustment'' of their actions to: (a) accommodate changes in the location of the target (as in test 1) (p. 7), or to (b) accommodate changes in the context of reality (such as the removal of the barrier in test 2) (p. 8). Note that in this new model goal-attribution is not a purely cue-based process any more, as it also entails the assumption of ''appropriate adjustment'' of means actions.

Admittedly, by this move P&P have increased the predictive power of their model significantly, however, we should note, only as a result of smuggling back the assumption of rationality of means action into their theory! We invite the reader to substitute ''rational'' for ''appropriate'' in P&P's formulations and see for herself/himself if there remains any difference between the predictions generated from our original proposals and from their follow-up version.

Maybe then P&P's criticism reduces to an argument about the semantic appropriateness of using the terms ''rational'' vs. ''appropriate'' in describing the adjustments required of goal-directed actions. The crucial issue then is what P&P mean by ''appropriate adjustment'' and whether their intended meaning can be demonstrated to differ significantly from our use of ''rational adjustment''. A healthy terminological debate is unlikely at this point, however, because just what makes a given adjustment of action ''appropriate'' is left largely unspecified in their model [apart from some vague intuitive suggestions concerning infants purported capacity to distinguish ''natural'' from unnatural motion (p. 7) and one example that ''if the target moves to the right, the object will not move to the left'' (p. 7)]. To make some headway, however, below we shall attempt to clarify *our* reasons for using the term 'rationality' in our theory of goal-attribution in infancy.

The teleological stance We accept that calling the nonmentalistic teleological interpretational system in one year-olds ''the intentional stance'' might have been misleading (in spite of our attempts to provide careful definitions). In fact, since then we have proposed (see Gergely and Csibra, 1996; Csibra et al., 1996) to distinguish this level of representing goal-directed action from the young child's later emerging mentalistic 'belief–desire psychology' or intentional stance by calling the former the 'teleological stance' (cf. Keil, 1994).

We believe that the teleological stance is a qualitatively different but developmentally related interpretational system that is the precursor of the young child's intentional stance. The two stances differ significantly in that: (a) teleological explanations are nonmentalistic as they make reference only to actual and future states of reality (situational constraints and goal-states) and so they do not require representing intentional mind states, and (b) they provide teleological rather than causal interpretations of behaviour. However, the two stances also share important features: both make reference to future states (goals vs. contents of desires), and to constraints on actions (in terms of actual reality vs. contents of

beliefs) as explanatory entities, and both rely on the same core inferential principle (the principle of rationality) when reasoning about action. Developmentally, teleological interpretations are transformed into causal mentalistic ones by 'mentalizing' the explanatory constructs of the teleological stance: i.e., by turning representations of actual reality constraints into 'beliefs' (which mentally represent such constraints), and representations of future goal states of reality into 'desires' (which mentally represent goal-states).

The principle of rational action We argued (Gergely et al., 1995, pp. 170–3) that for predicting actions from the intentional stance the mere attribution of beliefs and desires is not sufficient: it is only by relying on the assumption of rational action that an inference can be made as to which of the multiple possible means actions an agent will perform to realize his/her desire given the constraints of his/her beliefs about reality. When taking the intentional stance what counts as rational is evaluated over the domain of mental representations of actual or fictional reality (desires and beliefs) as a function of a set of background assumptions (about the agent's repertoire of means, his/her current state of resources, etc.). Depending on the richness of these factors, the computations identifying optimal action can get admirably complex in adults.

The point we wish to emphasize is that the teleological stance also needs an analogous principle to allow for the evaluation of multiple alternatives, if the stance is to be used for predicting or explaining action. This principle will play the *very same functional role* than that played by the assumption of rationality in the intentional stance. Of course, the domain of its application will be different (being actual and future states of reality rather than beliefs and desires) as will be the (much more restricted) set of background conditions (which may consist of ''the infant's knowledge of the physical, causal, gravitational and biomechanical constraints on the spatial movements of objects and agents'', see Gergely et al., 1995, pp. 1, 187). In fact, it may turn out that within the bounds of the teleological stance the notion of ''most rational spatial goal-approach'' will coincide with ''taking the shortest path'' (see Gergely and Csibra, 1996), though we believe this to be an empirical question. For P&P, however, ''to reduce rationality to: 'taking the shortest path''' (p. 10) in this manner does not do justice to the intricacies of the adult's mature notion of rationality.

In our view, however, it is not the concept of rationality that is reduced in the infant's teleological stance, rather, what is restricted is the domain over which it is applied and the set of background conditions over which it is computed. By insisting that it is the very same core principle of rational action of the intentional stance that is applied in teleological reasoning as well we want to focus attention on: (a) the functional isomorphism that this principle plays in the teleological and the intentional stance, and on (b) the developmental relatedness and continuity that we believe exists between the young child's theory of mind and its precursor, the teleological stance.

(3) Let us briefly reply to P&P's objection that our ''treatment of rationality... has the untenable consequence of completely eliminating the category 'irrational' or 'nonrational'. That is, if an object must be rational in order to be intentional, there can be no nonrational intentional objects'' (p. 10).

The misunderstanding here arises from a confusion about the identity of the object to which rationality is attributed. We believe that when interpreting behaviour as goal-directed, rationality is attributed as a property of the *action*, and not of the agent (or of the agent's mind). (Admittedly, we may have contributed to this confusion by indiscriminately referring in Gergely et al. (1995) to agents as well as actions of agents as being attributed rationality.) For an action to be represented as intentional it indeed seems necessary that it be evaluable as rational: a behaviour that cannot be justified in relation to the agent's beliefs and desires, will not be considered intentional.

In contrast, when we attribute rationality to an *agent*, we make a generalization over the whole range of his/her actions: if his/her actions can be generally construed as rational, we categorize the agent as a rational intentional person. Note that this allows for the case of intentional beings sometimes acting nonrationally. If we see George stumble (a behaviour we cannot interpret as a rational means action), we do not jump to the conclusion that George is irrational, we simply suspend the intentional stance in interpreting this particular behaviour of his. However, if we find that the majority of George's actions cannot be interpreted as rational, we may end up with the generalization that George as a person is not rational and as such cannot be held accountable for his actions.

Finally, P&P's further objection concerning irrational preferences in intentional agents seems to us to be a red herring precisely for the same reasons. The intentional stance does not judge the rationality of *goals* in relation to external (biological or social) standards, rather, it evaluates the rationality of an agent's particular *actions* (its proper domain) in relation to his/her beliefs and desires.

References

Csibra, G., Gergely, G., Bíró, S. and Koós, O. (1996). The perception of 'pure reason' in infancy, submitted.

Gergely, G. and Csibra, G. (1996). Understanding rational actions in infancy: Teleological interpretations without mental attribution. Paper presented at the *10th Biennial International Conference on Infant Studies*, April 19, 1996, Providence, Rhode Island, USA.

Gergely, G., Nádasdy, Z., Csibra, G. and Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165–193.

Keil, F.C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. Hirschfeld and S. Gelman, (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 234–254). New York: Cambridge University Press.

Premack, D. and Premack. A.J. (in press). The primacy of goals. *Cognition*.