

COOPERATION WITHOUT AWARENESS: A MULTIPERSON GENERALIZATION OF THE MINIMAL SOCIAL SITUATION*

by Andrew A. Coleman, Andrew M. Colman, and Richard M. Thomas

University of Leicester, England

The minimal social situation, which arises in living systems and subsystems at the level of the group, is a two-person game of incomplete information in which the players are ignorant of their interdependence. The win-stay, lose-change principle, based on the law of effect, explains how they nonetheless learn to cooperate when the game is repeated many times. In this paper the minimal social situation is generalized to groups of arbitrary size with the original two-person game representing a special case. Some theorems are derived from the assumption that the players follow the win-stay, lose-change principle, and the circumstances that result in joint cooperation are formally characterized. Whether or not an iterated multiperson minimal social situation results in joint cooperation under the win-stay, lose-change principle is shown to depend on the configuration of initial choices and the number of times that the group size is evenly divisible by two. Finally, some implications for experimental research are outlined.

KEY WORDS: individual organism, group, cooperation, decision making, game, minimal social situation.



INTRODUCTION

THE MINIMAL social situation, first described by Sidowski, Wyckoff, and Tabor (1956), is a two-person game of strategy in which the players are oblivious of their strategic interdependence. They are objectively interdependent because their payoffs are determined by each other's choices, but they are ignorant not only of the payoff structure of the game (as in other games of incomplete information) but also of the fact that they are involved in a game of strategy.

The following lifelike interpretation of the minimal social situation (Colman, 1982a, pp. 289–291) provides a useful intuitive background. Two commuters travel on the same train every day. They always sit in adjacent compartments, both of which are uncomfortably cold. Each compartment has a lever marked "heater," but there is no indication as to whether it should be turned to the left or right to increase the temperature. What the commuters do not know is that there is a fault

in the electrical wiring of the train: moving either lever to the left increases the temperature and moving either lever to the right decreases the temperature in the adjacent compartment. As a consequence of this, when either of the commuters turns the lever to the left or right, the other commuter is rewarded with warmth or punished with cold. The commuters cannot influence their own payoffs directly; their comfort or discomfort depends entirely on each other's choices, though neither of them realizes this. They would nonetheless both benefit if they turned their levers to the left at the beginning of every journey. The following interesting question arises: Can they learn to cooperate in this way in spite of their ignorance of their interdependence and even, perhaps, of each other's existence? If so, then people can learn to cooperate without any deliberate intention or awareness of the need for cooperation and without even knowing that they are involved in a social interaction.

Sidowski, Wyckoff, and Tabor (1956) and Sidowski (1957) provided experimental evidence showing that pairs of subjects can—and generally do—learn to cooperate in the minimal social situation, and this finding has been replicated many times (Kelley, Thibaut, Radloff & Mundy, 1962;

* Requests for reprints should be addressed to Andrew M. Colman, Department of Psychology, University of Leicester, Leicester LE1 7RH, England. We wish to thank Hilary Craig for her help and encouragement.

Rabinowitz, Kelley & Rosenblatt, 1966; Arickx & Van Avermaet, 1981). In the original experiments, a situation strategically equivalent to the commuters' dilemma was engineered as follows. Pairs of subjects were seated in separate rooms, unaware of each other's existence, and electrodes were attached to their bodies. Each subject faced an apparatus comprising a pair of buttons, which for convenience we shall label 0 and 1, and a digital display showing the cumulative total of points scored. Their instructions were to press one of the buttons on each trial, attempting always to maximize rewards (points) and to minimize punishments (shocks). The electrical wiring was arranged in such a way that on every trial a 0 choice delivered a point and a 1 choice a shock to the *other* subject. In more recent experiments similar devices have been used, except that negative payoffs have usually involved deduction of points rather than electric shocks.

The payoff matrix of the minimal social situation is shown in Figure 1. It is unnecessary to assign numerical values to the matrix elements: we need to assume only that each player prefers a positive payoff (+) to a negative payoff (-). One player chooses row 0 or 1, and the other player chooses column 0 or 1. If both choose 0, then the outcome is the top left cell and both receive positive payoffs. If both choose 1, then both receive negative payoffs. If one chooses 0 and the other chooses 1, then the 0-chooser receives a negative payoff and the 1-chooser receives a positive payoff. The rules of the game usually stipulate that the players choose simultaneously, although sequential choosing has also been studied. The payoff structure of Figure 1, called *mutual fate control* by Thibaut and Kelley (1959), has been used in most empirical investigations of the minimal social situation. It is obviously isomorphic with the commuters' dilemma described earlier.

When the game is repeated many times, pairs of subjects generally learn to coordinate their choices while remaining unaware of their strategic interdependence (Sidowski, Wyckoff & Tabor, 1956; Sidowski, 1957; Kelley, Thibaut, Radloff & Mundy, 1962; Rabinowitz, Kelley & Rosenblatt,

	0	1
0	+	+
1	-	-

FIG. 1. The "mutual fate control" payoff matrix of the minimal social situation. One player chooses between rows 0 and 1, and the other chooses between columns 0 and 1. The positive or negative payoff in the lower left half of each cell goes to the row-chooser, and the payoffs in the upper right halves go to the column-chooser.

1966; Arickx & Van Avermaet, 1981). Although they usually assume (incorrectly) that their payoffs are determined in some way by their own choices, they tend to choose 0 with increasing frequency over trials. In the long run, pairs of subjects often settle down to choosing 0 on every occasion. Subjects behave as if they were learning to cooperate, although from their point of view the situation is entirely non-social. How can this effect be explained? Kelley et al. proposed that subjects in the minimal social situation and similar games of incomplete information learn to adopt a *win-stay, lose-change* principle, which is merely an application of Thorndike's (1911) law of effect. The principle does not generate any prediction about the players' initial choices, but if the game is played more than once it implies that a player will repeat any choice that is followed by a positive payoff and switch to the other choice after receiving a negative payoff. If both players choose 0 on the first trial, for example, then both receive positive payoffs and, according to the win-stay, lose-change principle, both will choose 0 on the second

and all subsequent trials. We can represent the outcomes on successive trials by a sequence of ordered pairs corresponding to the row and column players' choices respectively:

$$(0, 0), (0, 0), (0, 0), \dots$$

If both players choose 1 on the first trial, then both receive negative payoffs which cause them to switch to 0 on the second trial, and these 0 choices are repeated on all subsequent trials:

$$(1, 1), (0, 0), (0, 0), (0, 0), \dots$$

If one player initially chooses 0 and the other chooses 1, then the 0-chooser receives a negative payoff and therefore switches to 1 on the second trial, and the 1-chooser receives a positive payoff and therefore sticks to 1 on the second trial. On the second trial, therefore, both players will choose 1, followed (as shown above) by 0 on all subsequent trials:

$$(0, 1), (1, 1), (0, 0), (0, 0), \dots$$

$$(1, 0), (1, 1), (0, 0), (0, 0), \dots$$

It is clear from this analysis that players who follow the win-stay, lose-change principle learn to cooperate—to choose mutually rewarding strategies—by the third trial at the latest, and continue to cooperate indefinitely after that.

Experimental evidence shows, however, that people do not generally follow the win-stay, lose-change principle rigidly (Rabinowitz, Kelley & Rosenblatt, 1966; Burnstein, 1969; Arickx & Van Avermaet, 1981; Colman, 1982b). In general, cooperative 0 choices begin to exceed chance frequency after a few trials and continue to increase in frequency; after 100 trials about 75% of choices are cooperative. According to the win-stay, lose-change principle, of course, 100% cooperation should occur after three trials. This means that players in the minimal social situation do not obey the law of effect strictly. But cooperative behavior does tend to evolve and, in the light of overwhelming evidence from other branches of psychology, it seems reasonable to assume that people are governed to a large extent by the law of effect and there-

fore that they *tend* to follow the win-stay, lose-change principle. This implies that the *probability* of a player's choice on trial t being repeated on trial $t + 1$ increases if the player is rewarded and decreases if the player is punished on trial t (Arickx & Van Avermaet, 1981).

In the sections that follow, we propose to generalize the minimal social situation to groups of arbitrary size. We shall then investigate the consequences of the win-stay, lose-change principle in these n -person games and characterize the circumstances that result in joint cooperation.

GENERALIZATION TO N -PERSON GROUPS

Preliminary formalization

The n -person minimal social situation is a game involving $n \geq 2$ players, each of whom has a uniquely designated *predecessor* and *successor*. The game can be represented by a cyclic graph of valency 2. It is useful to imagine the n players sitting round a table, so that 1's predecessor is n and n 's successor is 1. Each player has a choice of two strategies, 0 and 1, so the choices of the n players on a specified trial can be represented by an n -vector of zeros and ones which we call a *configuration*. If a player chooses 0, then that player's successor receives a positive payoff, and if a player chooses 1, then that player's successor receives a negative payoff. According to the win-stay, lose-change principle, any player who receives a positive payoff will repeat the same strategy choice on the following trial, and any player who receives a negative payoff will switch strategies on the following trial. For any configuration, therefore, there is a unique configuration that follows it according to the win-stay, lose-change principle. A configuration consisting entirely of zeros will be repeated on all subsequent trials. Any configuration that leads ultimately to this zero configuration is called *cooperative*. The analysis in the previous section shows that in the two-person minimal social situation, which is merely a special case of the general n -person game, all configurations are cooperative.

Some typical configurations in a six-person minimal social situation will illustrate these ideas. In this game, the configuration (1, 0, 1, 0, 1, 0) is followed on the next trial by (1, 1, 1, 1, 1, 1), and then by (0, 0, 0, 0, 0, 0); the initial configuration (1, 0, 1, 0, 1, 0) is therefore cooperative. On the other hand, the configuration (1, 0, 1, 0, 0, 0) generates the following sequence: (1, 0, 1, 0, 0, 0), (1, 1, 1, 1, 0, 0), (1, 0, 0, 0, 1, 0), (1, 1, 0, 0, 1, 1), (0, 0, 1, 0, 1, 0), (0, 0, 1, 1, 1, 1), (1, 0, 1, 0, 0, 0), and the initial configuration is repeated. Such a configuration will evidently cycle for ever, never reaching (0, 0, 0, 0, 0, 0), which shows that the initial configuration is noncooperative.

An arbitrary configuration in an n -person minimal social situation can be represented by the vector

$$(x_1, x_2, \dots, x_n),$$

where $x_i \in \{0, 1\}$. The numbers 0 and 1 can be regarded as elements of the field $GF(2)$ of integers modulo 2. In the configuration (y_1, \dots, y_n) immediately following (x_1, \dots, x_n) , since $0 + 0 = 1 + 1 = 0$, and $1 + 0 = 0 + 1 = 1$ in $GF(2)$,

$$y_i = \begin{cases} x_i & \text{if } x_{i-1} = 0 \\ x_i + 1 & \text{if } x_{i-1} = 1, \end{cases}$$

where the subscripts are reduced modulo n . Therefore,

$$y_i = x_{i-1} + x_i \quad (i = 1, \dots, n).$$

The configuration immediately following (x_1, \dots, x_n) is therefore obtained by applying the linear transformation

$$T: (x_1, \dots, x_n)' \rightarrow (x_n + x_1, x_1 + x_2, \dots, x_{n-1} + x_n)'$$

where x' denotes the transpose of the row vector x . The transformation matrix is an n -square matrix $T = [t_{ij}]$ in which

$$t_{ij} = \begin{cases} 1 & \text{if } i = j \text{ or } i = j + 1 \\ 0 & \text{otherwise.} \end{cases}$$

The general form of the transformation matrix is shown in Table 1.

If the initial configuration is $x = (x_1, \dots, x_n)$, then the sequence of configurations (represented by transposed row vectors) on subsequent trials will be Tx' , T^2x' , T^3x' ,

TABLE 1
TRANSFORMATION MATRIX FOR THE N -PERSON
MINIMAL SOCIAL SITUATION UNDER THE WIN-STAY,
LOSE-CHANGE PRINCIPLE.

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \dots & \dots & \dots & 1 \\ 1 & 1 & 0 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & 1 & 1 & 0 & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1 & 1 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \dots & 0 & 1 & 1 \end{pmatrix}$$

... An initial configuration is cooperative, therefore, if $T^k x' = (0, 0, \dots, 0)'$ for some k , that is, if $x = (x_1, \dots, x_n)$ lies in the kernel of the linear transformation T^k for some k .

FORMAL RESULTS

Theorem 1. If the configuration (x_1, \dots, x_n) is followed immediately by $(0, 0, \dots, 0)$, then either $(x_1, \dots, x_n) = (0, 0, \dots, 0)$ or $(x_1, \dots, x_n) = (1, 1, \dots, 1)$.

Proof. If $x_i = 0$, then $x_{i-1} = 0$, otherwise the i th component of the transformed vector Tx' would be 1. Similarly, bearing in mind that $1 + 1 = 0$ in $GF(2)$, if $x_i = 1$, then $x_{i-1} = 1$, otherwise the i th component of the transformed vector would be 1. This theorem establishes that the only configurations that are immediately followed by joint cooperation are those in which all players make the same choice.

Theorem 2. If n is odd, then the only cooperative configurations are $(0, 0, \dots, 0)$ and $(1, 1, \dots, 1)$.

Proof. If $Tx' = (0, 0, \dots, 0)'$, then it follows from Theorem 1 that either $x = (0, 0, \dots, 0)$ or $x = (1, 1, \dots, 1)$, and if $(0, 0, \dots, 0)$ is not the initial configuration, then it must be preceded by $(1, 1, \dots, 1)$. Suppose that $(1, 1, \dots, 1)$ is also not the initial configuration. Then $Tw' = (1, 1, \dots, 1)'$ for some w . Now if $w_i = 0$, then $w_{i-1} = 1$, otherwise the i th component of Tw' would be zero. For the same reason, if $w_i = 1$, then $w_{i-1} = 0$. Therefore, $w_{i-2} = w_i$. Consider the vector component w_n : since n is odd, $w_n = w_{n-2} = \dots = w_1$. This implies that if $w_1 = 0$, then $w_{1-1} = w_n = 0$, and if $w_1 = 1$, then $w_{1-1} = w_n = 1$, which yields a contradiction. We have therefore proved that if the num-

ber of players is odd, joint cooperation is achieved only if all players make the same initial choice, and it results after one trial at most.

Theorem 3. If $j = 2^p$, $p \in Z^+$ (where Z^+ is the set of nonnegative integers), then

$$T_j: (x_1, \dots, x_n) \rightarrow (x_{1-j} + x_1, \dots, x_{n-j} + x_n),$$

where the subscripts are expressed modulo n .

Proof. Assume that the result is true for some p . Then, if $q = 2^p$,

$$T^q: (x_1, \dots, x_n) \rightarrow (x_{1-2q} + x_1, \dots, x_{n-2q} + x_n).$$

The proof proceeds by induction on p . For $p + 1$, $2^{p+1} = 2q$, and $T^{2q} = T^q T^q$. Now

$$T^{2q} = T^q T^q: (x_1, \dots, x_n) \rightarrow (y_1, \dots, y_n),$$

where

$$y_i = (x_{i-2q} + x_{i-q}) + (x_{i-q} + x_i) = x_{i-2q} + x_i$$

because, whether $x_{i-q} = 0$ or 1 , $x_{i-q} + x_{i-q} = 0$. Thus,

$$T^{2q}: (x_1, \dots, x_n) \rightarrow (x_{1-2q} + x_1, \dots, x_{n-2q} + x_n).$$

We have proved that if the result holds for some p then it holds for $p + 1$. The final step is to show that it holds for $p = 0$. In that case $q = 2^0 = 1$, and $T^1 = T$ is the basic transformation

$$T: (x_1, \dots, x_n) \rightarrow (x_n + x_1, x_1 + x_2, \dots, x_{n-1} + x_n)$$

for which the result holds. We have therefore proved that any number of trials j that is a power of 2 takes each component of x_i of the configuration x into $x_{i-j} + x_i$.

Theorem 4. A configuration (x_1, \dots, x_n) in an n -person minimal social situation is cooperative iff $x_i = x_{i-k}$ for all i , where $n = bk$, $k = 2^a$, $a, b, \in Z^+$ (the set of nonnegative integers), and b is odd.

Proof. From Theorem 3 we have

$$T^j: (x_1, \dots, x_n) \rightarrow (x_{1-j} + x_1, \dots, x_{n-j} + x_n),$$

where $j = 2^p$, $p \in Z^+$. The kernel of T^j is therefore

$$\begin{aligned} \ker T^j &= \{(x_1, \dots, x_n) \mid x_{1-j} + x_1 = \dots = x_{n-j} + x_n = 0\} \\ &= \{(x_1, \dots, x_n) \mid x_i = x_{i+j} \text{ for all } i\}. \end{aligned}$$

Since $\ker T^p$ is a subset of $\ker T^{p+1}$ for all natural numbers p , the set of cooperative states is $\ker T^p$ if $\ker T^p = \ker T^m$ for $p < m$. The proof is constructive: we shall prove that if $k = 2^a$, $m = 2k = 2^{a+1}$, then $\ker T^k = \ker T^m = \ker T^{2k}$.

Let $c = (b + 1)/2$. Then $b = 2c - 1$, and hence $kb = k(2c - 1)$. Thus $2ck \equiv k \pmod{kb}$, that is,

$$cm \equiv k \pmod{n}.$$

Now, if $x \in \ker T^m$, then $x_i = x_{i+m}$ for all $i \pmod{n}$. It follows that $x_{i+m} = x_{i+2m} = x_{i+3m} \dots$, and therefore, since c is a positive integer, that

$$x_i = x_{i+cm} \text{ for all } i \pmod{n}.$$

Since $cm \equiv k \pmod{n}$,

$$x_i = x_{i+k} \text{ for all } i \pmod{n},$$

which shows that $x \in \ker T^k$. We have therefore proved that $\ker T^m$ is a subset of $\ker T^k$, and hence, since $k < m$, that

$$\ker T^k = \ker T^m,$$

as required.

This theorem shows that we can characterize the cooperative configurations in an n -person minimal social situation as follows. If n is odd, then the cooperative configurations are (x_1, \dots, x_n) such that $x_i = x_{i+1}$ for all $i \pmod{n}$. If n is even, then if k is the highest power of 2 that divides n evenly, then the cooperative configurations are (x_1, \dots, x_n) such that $x_i = x_{i+k}$ for all $i \pmod{n}$.

The proof also implies that, in an n -person minimal social situation, if k is the highest power of 2 that divides n evenly, the first k players may choose arbitrarily, but the choices of the remaining players

are determined for the configuration to be cooperative. It follows that the number of cooperative initial configurations is 2^k .

DISCUSSION

An abstract theory, if it is to be useful, should do two things. First, it should explain existing empirical data. Second, it should provide conclusions whose scope extends beyond existing data but can be empirically tested. The formal system developed in this paper explains existing data in a trivial sense, inasmuch as it incorporates the theory and experimental findings related to the two-person minimal social situation as a special case.

The win-stay, lose-change principle, which has been used to explain the evolution of cooperation in the two-person game, is derived from the law of effect, originally formulated by Thorndike (1911) as follows: "Responses . . . which are accompanied or closely followed by satisfaction [are] more firmly connected with the situation . . . ; those which are accompanied or closely followed by discomfort . . . have their connections with the situation weakened" (p. 244). (Thorndike later "repealed" the second part, which is sometimes called the negative law of effect.) Many behaviorist psychologists, including Skinner (1966, 1984), regard the law of effect as a behavioral parallel of natural selection in which only the most successful responses in an organism's behavioral repertoire survive while the unsuccessful responses become extinct. Several decades of psychological research have provided abundant corroboration of the law of effect in a wide variety of situations. It would be most surprising if it were found not to apply in the minimal social situation. It is worth pointing out, however, that the win-stay, lose-change principle is an idealized version of the law of effect in which responses that are rewarded are *invariably* repeated and those that are punished are *never* repeated on the following trial—the principle, unlike the law of effect, is deterministic and noncumulative. It is clear, however, that the broad outline of the theory presented in this paper would remain valid if a stochastic version of the win-stay,

lose-change principle based on probabilistic learning theory were substituted.

The formal results certainly extend beyond the scope of existing empirical data. Many of the predictions that can be derived from the analysis are counterintuitive but nevertheless easily testable. Among the interesting predictions that should be tested are the following. First, although the frequency of rewarding choices and joint cooperation tends to increase in the two-person minimal social situation, the theory predicts no such increase in odd-sized groups. Second, whenever the number of players is even but not a power of two, configurations that are cooperative according to the theory should progress toward joint cooperation more frequently than noncooperative configurations. Third, multiperson groups in which the number of players is a power of two should behave like players in the two-person minimal social situation: irrespective of the choices made on the first trial, there should be steady progress toward joint cooperation. Fourth, the frequency of rewarding choices and joint cooperation should correlate with the number of cooperative initial configurations determined by the theory. If any of these predictions turns out to be wrong, then the assumptions of the theory will have to be modified.

REFERENCES

- Arickx, M. & Van Avermaet, E. Interdependent learning in a minimal social situation. *Behavioral Science*, 1981, 26, 229–242.
- Burnstein, E. The role of reward and punishment in the development of behavioral interdependence. In J. Mills (Ed.), *Experimental social psychology*. London: Macmillan, 1969, 341–405.
- Colman, A. M. Conclusions. In A. M. Colman (Ed.), *Cooperation and competition in humans and animals*. Wokingham: Van Nostrand Reinhold, 1982a, 285–294.
- Colman, A. M. *Game theory and experimental games: The study of strategic interaction*. Oxford: Pergamon, 1982b.
- Kelley, H. H., Thibaut, J. W., Radloff, R. & Mundy, D. The development of cooperation in the "minimal social situation." *Psychological Monographs*, 1962, 76, Whole No. 19.
- Rabinowitz, L., Kelley, H. H. & Rosenblatt, R. M. Effects of different types of interdependence and response conditions in the minimal social situation. *Journal of Experimental Social Psychology*, 1966, 2, 169–197.

- Sidowski, J. B. Reward and punishment in the minimal social situation. *Journal of Experimental Psychology*, 1957, 54, 318-326.
- Sidowski, J. B., Wyckoff, L. B. & Tabor, L. The influence of reinforcement and punishment in a minimal social situation. *Journal of Abnormal and Social Psychology*, 1956, 52, 115-119.
- Skinner, B. F. The phylogeny and ontogeny of behavior. *Science*, 1966, 153, 1205-1213.
- Skinner, B. F. The phylogeny and ontogeny of behavior [updated, with peer commentary]. *The Behavioral and Brain Sciences*, 1984, 7, 669-711.
- Thibaut, J. W. & Kelley, H. H. *The social psychology of groups*. New York: Wiley, 1959.
- Thorndike, R. L. *Animal intelligence: Experimental studies*. New York: Macmillan, 1911.

(Manuscript received May 16, 1989.)