# A Committee and a Puzzle of Costs and Benefits

Athanassios Pitsoulis*

February 18, 2008

**Abstract**

We explore a common knowledge puzzle applied to a deliberative assembly where factions have heterogeneous preferences and private information regarding the true state of the world. It is shown that the members do not need to communicate directly but can infer the true state of the world from a vote on whether everybody shall disclose their private information. Commitment to disclosure serves as a signal regarding knowledge and makes direct communication unnecessary.

**JEL classification** C79, D82

**Keywords** Common knowledge, Puzzle of the Hats

## 1 Introduction

The Puzzle of the Hats, first mentioned by Littlewood (1953), is an old puzzle depending on reasoning about the reasoning of others (Geanakoplos 1994). It can be told in many equivalent ways and is also known as the Dirty Faces Game. The puzzle is about $n$ "logical" individuals wearing a hat that is either white or black.[1] Each individual knows that everybody is wearing a hat and can see all hats except the own. An observer asks those individuals who know the color of the own hat to raise a hand but no one raises a hand no matter how often the observer asks. Then the observer announces that at least one

*Institute of Economic Sciences, Chair of Microeconomics, Brandenburg University of Technology, Konrad-Wachsmann-Allee 1, D-03046 Cottbus, Germany. Phone: +49 355 69 2982, Fax: +49 355 3020, e-Mail: pitsouli@tu-cottbus.de.

[1]In the Dirty Faces version the individuals are children with either clean or dirty faces.

of the hats is white and goes on asking the individuals to raise a hand if the color of the own hat is known. If $k$ hats are white no one raises a hand until the $k$-th question, at which suddenly all individuals with white hats do so.

The apparently ignorant individuals learned by making inferences about each other by using common knowledge. This important function of common knowledge was emphasized by Schelling (1960) and Harsanyi (1967,8), the first explicit analysis was given by Lewis (1969). This paper explores a new variant of the Puzzle of the Hats applied to a deliberative assembly where members have heterogeneous preferences and private information regarding the true benefits and costs of a project. One would expect direct communication but the agents can infer the true state of the world from a vote on whether the committee's members would commit to disclosure of their private information. The result of this vote solves the puzzle of benefits and costs.

# 2 The Problem

We look at a deliberative assembly ("the committee") consisting of individuals $1, 2, \ldots k, \ldots, K$ where $K \in \mathbb{N}^*$. The committee is to decide by simple majority whether nor not to realize a project (for instance whether or not to start a war, prohibit smoking, build a dam etc.) We assume the committee is presided by a nonpartisan, disinterested chairperson. The members' voting intentions ("yea" or "nay") are depending on their preferences and the state of the world. The state of the world is reflected in two variables, per-capita benefits $b$ and costs $c$.[2] Four states of the world are possible: The project can bring low benefits and low costs, $(b_l, c_l)$, high benefits and low costs, $(b_h, c_l)$, low benefits and high costs, $(b_l, c_h)$, or high benefits and high costs, $(b_h, c_h)$.

We start from a given initial distribution of individuals belonging to four types of individual preferences: A committee member $k$ can either a right-wing hawk ($R_H$), a right-wing dove ($R_D$), a left-wing hawk ($L_H$) or a left-wing dove ($L_D$). Right-wing hawks always vote for the project, regardless of costs and benefits. Right-wing doves always vote for the project unless the benefits are low and the costs high. Left-wing doves always vote against the project unless the benefits are high and the costs low. Left-wing hawks, somewhat inconsistently, vote for the project if the benefits are high and the costs low, or the benefits are low and the costs high, otherwise they vote against the project. The preference-types are shown in Table 1. This table is common

[2] An alternative setting is that the committee members are to elect a candidate by majority vote. Instead of benefits and costs the information received could reflect the candidate's policy proposals or personal characteristics.

knowledge among all committee members. The distribution of preference types is unknown.

Table 1: Preference Types

| | $c_h$ | | $c_l$ | |
|---|---|---|---|---|
| $b_h$ | $R_H$: Y | $R_D$: Y | $R_H$: Y | $R_D$: Y |
| | $L_H$: N | $L_D$: N | $L_H$: Y | $L_D$: Y |
| $b_l$ | $R_H$: Y | $R_D$: N | $R_H$: Y | $R_D$: Y |
| | $L_H$: Y | $L_D$: N | $L_H$: N | $L_D$: N |

Y: "Yea", N: "Nay"

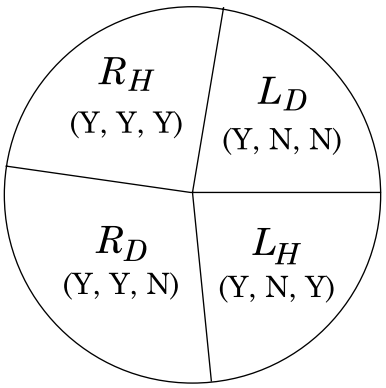
At the beginning of the game nature reveals a *public* cost-benefit-signal, i.e. one received by every individual. We assume this signal to be the information $(b_h, c_l)$. This means that every member of the committee would vote *for* the project at this stage. In the next phase new information becomes available but nature this time reveals only a *private* signal containing some information about the true state of the world. This fact is common knowledge among the committee members, but the content of the signal is not. We assume the true state of the world at this stage to be $(b_l^*, c_h^*)$, with the asterisk denoting correct information, and that the right-wing types receive the signal $b_l^*$ while left-wing types receive the signal $c_h^*$. It is assumed that the individuals do not know that nature *either* sends a new benefit *or* cost signal but that some individuals may have received *both* a new benefit and cost signal.

This assumption is crucial for the following reason. Suppose the committee members would know that either a new cost or benefit signal were received. A test poll at this stage would reveal to every individual the true state of the world. Suppose the initial test vote reveals a majority for the "yea"-faction. A "yea"-voter, having received the signal $b_l^*$, can infer that the "nay"-voters must have received $c_h^*$ and vice versa. But suppose now the committee members can not be sure whether the other faction received not one signal but both. In that case a test vote does not reveal the new state of the world. A "yea"-voter, having received the signal $b_l^*$, can not infer that the "nay"-voters must have received $c_h^*$ as the "nay"-vote would be consistent with the signal $(b_l^*, c_l^*)$. $b_l^*$ would moreover confirm prior information.

It is natural that all right-wing members form one faction and all left-wing members another as the former would initially vote for the project and the latter against. We may call right-wing hawks as well as left-wing doves

'hardliners' and right-wing doves as well as left-wing hawks 'floating voters'. In a static world costless and honest mutual disclosure of private information would aggregate the committee members' decentralized knowledge until everybody is perfectly informed about the true state of the world. The result would be that all hawks vote for and all doves against the project. But is it necessary that information is disclosed in the committee to get this result? The answer is: no!

Figure 1: The Committee




(Y,Y,N) reads: Votes "yea" with initial information, "yea" after having received new information, "nay" with complete information

# 3 Solution

Disclosure of private information seems to be the intuitive solution to the problem, but it can be shown that it is not necessary in the committee. The disclosure of the willingness to disclose indirectly communicates the information needed for other agents to infer the true state of the world.

Suppose the committee's chairperson demands a poll on whether the members commit to disclose private information before the final vote on the project. The incentives to disclose depend on which faction has a majority. Suppose the test vote revealed to the individuals a majority of "yea"-votes. Right-wing hawks would prefer to vote immediately but the left-wing doves would have an incentive to disclose the high costs. Knowing this the right-wing hawks were better off by committing to full disclosure, too. But what about the floating voters?

A right-wing dove knows that the left-wingers either received the signal $(b_l^*, c_l^*)$ or $c_h^*$. In the first case she would rightly be voting for the project

and disclosing private information would not change the outcome. In the second case her disclosure of the signal $b_l^*$ to a left-wing hawk would induce the latter to vote *for* the project. In other words: The right-wing dove knows that she might be wrong about the costs. She knows, too, that she might be right without knowing it. From her point of view it were better if only the left-wingers would disclose their private information. The same holds for the other side. Floating voters are obviously caught in the dilemma that if they are mistaken about the true state of the world so may be their counterparts on the other side. They thus have no incentive to commit to the disclosure of private information.

Interestingly, the poll on full disclosure provides the information needed in order to infer the true state of the world: The vote reveals that on both sides there are individuals who do *not* want to commit to disclosure. This supports the other side's inference that unfavorable information is being withheld and thereby reveals the other side is not sure about the true state of the world. This, finally, reveals the true state of the world $(b_l^*, c_h^*)$.

More formally, let the set of possible states of nature $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ represent the uncertainty which an agent faces when making a decision.[3] Here we have three possible worlds, corresponding to the following states of affairs:

$\omega_1$: Neither faction knows both the true costs and benefits $\Leftrightarrow$ Floating voters do not commit to disclosure

$\omega_2$: $L$-types know both the true costs and benefits $\Leftrightarrow$ $L_H$-types commit to disclosure

$\omega_3$: $R$-types know both the true costs and benefits $\Leftrightarrow$ $R_D$-types commit to disclosure

The knowledge set on an agent of the type $i$ $(i = R, L)$ is represented by an information partition $\mathcal{H}_i$ of the set $\Omega$.

**Definition 1** An information partition $\mathcal{H}_i$ is a collection $\{h_i(\omega)|\omega \in \Omega\}$ of disjoint subsets of $\Omega$ such that (a) $\omega \in h_i(\omega)$ and (b) if $\omega' \in h_i(\omega)$ then $h_i(\omega') = h_i(\omega)$.

Here we have $\mathcal{H}_R = \{\{\omega_1, \omega_2\}, \{\omega_3\}\}$ and $\mathcal{H}_L = \{\{\omega_1, \omega_3\}, \{\omega_2\}\}$. So a $R$-type individual knows that either $\{\omega_1, \omega_2\}$ or $\{\omega_3\}$ is the case, while a $L$-type individual knows that either $\{\omega_1, \omega_3\}$ or $\{\omega_2\}$ is the case.

We now introduce the knowledge functions.

[3] This is based on the formalization in Osborne and Rubinstein 1994: 67-85.

**Definition 2** For any event $E$ we have $K_i(E) = \{\omega \in \Omega | h_i(\omega) \subseteq E\}$.

After the test vote and before the vote on disclosure we have $\mathcal{K}_R = \emptyset$ and $\mathcal{K}_L = \emptyset$. This means neither of both types knows for sure whether $\omega_1$, $\omega_2$ or $\omega_3$ is true. With a slight abuse of notation we can write these expressions as $\tilde{\mathcal{K}}_R = \{(b_l^*, c_l)\}$ and $\tilde{\mathcal{K}}_L = \{(b_h, c_h^*)\}$.

What happens if the committee members are asked whether or not they commit themselves to disclose their private information? The decision-problem of the types $R_D$ and $L_H$ can be formulated by help of the second-order mutual knowledge functions $\tilde{\mathcal{K}}_i(\tilde{\mathcal{K}}_j)$ $(i, j = R, L;\ i \neq j)$. Under the assumptions made here $\tilde{\mathcal{K}}_R(\tilde{\mathcal{K}}_L) = \{(b_h, c_h^*), (b_l^*, c_l^*)\}$ and $\tilde{\mathcal{K}}_L(\tilde{\mathcal{K}}_R) = \{(b_h^*, c_h^*), (b_l^*, c_l)\}$. The decision-problem is shown in Table 2.

Table 2: To Disclose or Not: Second-Order Decision-Problem of the Floating Voters

| $R_D$ \ $L_H$ | $(b_h, c_h^*)$ | $(b_l^*, c_l^*)$ |
|---|---|---|
| $(b_l^*, c_l)$ | N,N | Y,Y |
| $(b_h^*, c_h^*)$ | Y,Y | — |

Y: "Yea", N: "Nay"

We next use a definition of common knowledge given in Osborne and Rubinstein (1994, Definition 73.2) to solve the game explicitly.

**Definition 3** An event $F \subseteq \Omega$ is self-evident between the agents if for all $\omega \in F$ we have $h_i(\omega) \subseteq F$ for all agents. An event $E \subseteq \Omega$ is common knowledge between all agents in the state $\omega \in \Omega$ if there is a self-evident event $F$ for which $\omega \in F \subseteq E$.

Clearly, the event "some agents are not committed to disclosure" makes $\omega_1$ self-evident. All agents update their information partition and exclude all states of nature where disclosure should take place (the states $\omega_2$ and $\omega_3$). All agents now know for sure the state of nature is $\omega_1$ and thus $\tilde{\mathcal{K}}_i = \{(b_l^*, c_h^*)\}$. They do not need to disclose private information and can immediately vote on the project. This is where the game ends.

# 4 Summary

We explored a new variant of the Puzzle of the Hats applied to a deliberative assembly where factions have heterogeneous preferences and private information. The members do however not need to deliberate but can infer the true

state of the world from a vote on whether everybody is committed to disclose their private information. The willingness to disclose serves as a signal regarding knowledge which is indirectly communicated via a vote. This leads to revelation of the true state of the world.

This game can be cast into an experimental setting in order to test the proposition that perfectly rational individuals should not disclose private information as indirect communication reveals that unfavorable information is withheld. It could serve as a classroom experiment, too.

# References


[1] Geanakoplos, J., 1994, Common Knowledge, in: R. Aumann and S. Hart, eds., Handbook of Game Theory, Vol. 2 (Elsevier Science, Amsterdam) 1438-1496.

[2] Harsanyi, J., 1967, Games with incomplete information played by "Bayesian" players, I: The basic model, Management Science 14, 159-182.

[3] Lewis, D., 1969, Convention: A Philosophical Study. (Harvard University Press, Cambridge).

[4] Littlewood, J.E., 1953, A Mathematician's Miscellany. (Methuen and Company Limited, London).

[5] Osborne, M.J. and A. Rubinstein, 1994, A Course in Game Theory. (MIT Press, Cambridge).

[6] Schelling, T., 1960, The Strategy of Conflict. (Harvard University Press, Cambridge).