Chapter 2

Overcoming Moral Hypocrisy in a Virtual Society

Markus Christen

Institute of Biomedical Ethics, University of Zurich, Pestalozzistr. 24, 8032 Zurich, Switzerland

christen@ethik.uzh.ch

Complying with moral norms increases the reputation of individuals in a society but imposes costs, or missed gains, upon the individual when confronted with temptations. Thus, moral hypocrisy—avoiding the cost of being moral while maintaining moral reputation—may be the optimal behavior of individuals, which is corroborated by psychological research and everyday experience. In this study, the effect of various social strategies—e.g., avoiding wrongdoers or disclosing hypocrites—on the prevalence of moral hypocrisy in a virtual society is evaluated. We show that diversity with respect to population and temptation density is crucial to overcome moral hypocrisy.

2.1 Introduction

Morality is a fundamental aspect of the societal organization of mankind. Standards of morality such as norms, values, and virtues

Complexity and the Human Experience: Modeling Complexity in the Humanities and Social Sciences
Edited by Paul Youngman, Mirsad Hadzikadic, and Theodore D. Carmichael

ISBN 978-981-4463-26-3 (Hardcover), 978-981-4463-27-0 (eBook)
www.panstanford.com

pervade the practical life of humans and safeguard the society from behaviors profitable for individuals but damaging for the group. Individuals fulfilling these standards gain a high reputation. This factor is considered to be an essential component for the development of morality in hunter-gatherer societies, where each individual is strongly aware that he or she must have a positive reputation in case of future need, and painfully guards it [11]. However, moral behavior can also involve disadvantages for an individual—either direct in the sense of "missed opportunities" or indirect in the form of costs that result from punishing wrongdoers. Concerning the latter, research shows that individuals seem to be willing to pay the costs of enforcing moral norms, e.g., through altruistic punishment [7]. Furthermore, the propagation of this strategy seems to increase with size [12] and social complexity [15] of a society. However, the fact that punishment is enforced also indicates that the temptation to trespass moral norms still exists, because the violator has gains—in particular, if the violation is not detected by anyone, e.g., the case of subtle cheating [18].

Thus, a tension between moral reputation and moral action emerges: The former has a beneficial effect for the individual, as he or she becomes a respected member of the society. However, fulfilling moral standards when confronted with specific temptations imposes losses on the individual that he or she may try to avoid. This leads to moral hypocrisy: avoiding the cost of being moral while maintaining the appearance of morality. There is no reliable data on the prevalence of this behavior in different kinds of societal organization, but it can be expected that moral hypocrisy is a widespread phenomenon in modern societies where social control is less effective compared to small-scale societies [4] and where the "opportunity space" increases due to modern technologies such as the Internet [3]; dating Web sites allowing married persons to find additional sexual partners are apt examples. The potentially substantial prevalence of moral hypocrisy is corroborated by social psychology research [1,8,17] and by the everyday observation that exemplars of moral hypocrisy—e.g., if they concern sexual moral norms like adultery—attract a great deal of attention in popular media.

In this study, we analyze moral hypocrisy from a modeling point of view by focusing on social strategies—examples are avoiding or disclosing hypocrites—that are integrated in a virtual

society aiming to counteract moral hypocrisy. This approach complements the current discussion on moral hypocrisy in social psychology that focuses either on the psychological factors of how individuals maintain a motivational state with the ultimate goal of appearing moral while, if possible, avoiding the costs to oneself of actually being moral [2], or why individuals' evaluations of their own moral transgressions may differ substantially from their evaluations of the same transgressions committed by others [19]. The next section presents the model and its validation. The results section illustrates the effect of various combinations of social strategies on the prevalence of moral hypocrisy for four paradigmatic scenarios. The model implements these scenarios both individually, called non-diversity condition, and as combination, called diversity condition. The concluding section contains a discussion of the relevance of the results obtained for potential real-world strategies against moral hypocrisy.

2.2 The Moral Hypocrisy Model

2.2.1 Conceptualization of Moral Hypocrisy

Our model implements the conceptual idea of moral hypocrisy by distinguishing two different types of agent-states: the reputation of the agent, either morally good (G) or bad (B), and its disposition to act toward temptations, that is, either to be tempted (T) or to resist a temptation (R). The combinations of these states offer four different behaviors to the agents: appearing good and resisting a temptation (GR; "good guys"), appearing good but being tempted (GT; "hypocrites"), appearing bad and being tempted (BT; "bad guys") and appearing bad but resisting temptations (BR; "inconsistent guys"). The model is spatial; the agents interact by comparing payoffs within their Moore neighborhood—the eight cells surrounding a central cell occupied by the agent on a two-dimensional square lattice—and follow, if allowed to do so, social strategies that may dislocate the agent on the lattice.

The payoff structure (see Table 2.1) depicts the advantage of moral hypocrisy, i.e., an agent gains most if it is tempted although its reputation is good. In other words, the model *assumes* that moral hypocrisy is the optimal behavior for a single agent within a

society. Therefore, it can be expected that "hypocrites" (GT agents) dominate within this virtual society unless other factors overcome this dominance. These factors can either be generic, e.g., population or temptation density that are predefined in the beginning. Or they can be interventional, i.e., they consist of a social strategy that intends to overcome the dominance of hypocrisy.

Table 2.1 Payoff-matrix for an agent that implements one of the four behavior types of the model

		Disposition to act	
		Be tempted (T)	**Resist temptation (R)**
Reputation	Good (G)	GT (yellow): +1 for each neighbor and each temptation in Moore Neighborhood	GR (blue): +1 for each neighbor in Moore Neighborhood
	Bad (B)	BT (red): +1 for each temptation in Moore Neighborhood	BR (pink): 0

Note: The color code refers to the figures (the online version has color graphics).

The goal of this study is to assess the success of different strategies compared to a benchmark, characterized by no strategy installed, in dependence of generic factors that model social complexity through a predefined population and temptation density. Success is measured in terms of changes in the population distribution of agents that follow one of the four behaviors. The model implements a basic conceptualization of moral hypocrisy, disregarding the specific type of temptation and internal psychological mechanisms that may, for example, explain why "inconsistent" behavior could be possible.

2.2.2 Model and Social Strategies Implementation

We implemented the moral hypocrisy model in NetLogo [20] using a 41 × 41 lattice that offers space for maximally 1681 agents and an equal number of temptations. There are four model parameters: two scenario and two initialization parameters. Scenario parameters are the agent density d_a and the temptation density d_t; initialization parameters are the probability p_r that an agent has a good reputation—otherwise, it has a bad reputation—and the probability

p_t that an agent is tempted, when a temptation is present in its Moore neighborhood—otherwise, it is not tempted.

There are two versions of the model: In the non-diversity version, both agents and temptations are randomly distributed on the lattice. Also in the diversity version, agents and temptations are spread randomly, but agent and temptation densities are different in the four quadrants of the lattice such that the four main scenarios (see Section 2.2.3) were simultaneously present in the model.

In both versions of the moral hypocrisy model, one of the four behaviors depicted in Section 2.2.1 is assigned to each individual agents based on predefined p_r and p_t. After initializing the model in this way, each agent obtains its payoffs based on the payoff-matrix (Table 2.1) as follows: Reputation is a function of the size of the agent's community, and "good" agents obtain as many points as there are other agents in its Moore neighborhood. The same holds for temptations, i.e., an agent that is disposed to be tempted will obtain as many points as there are temptations in its Moore neighborhood. For example, a GT agent with 3 agents and 2 temptations in its neighborhood will obtain five points in an update cycle. After updating where all agents are called in a random order, each agent compares its payoff with the payoff of its neighbors. If one neighbor has a higher payoff, the agent will adopt the behavior of the "winner"; e.g., turns from GR to GT. If not, the agent keeps its behavior; in case of tie, the agent switches its behavior with probability 0.5. After updating the behavior of all agents in this way, the sizes of the populations representing the four behavior types (GR, GT, BR, and BT) are counted and the next cycle begins.[a] If the change of the GT population in a consecutive cycle is smaller than 1% of the mean of the GT populations of the 10 previous cycles, the simulation has reached a quasi-stable state and stops.[b]

To make the result independent from the specific initialization with respect to the distributions of behaviors at the starting point,

[a]Updating with respect to payoff calculation and following of the implemented strategy have been checked by hand in a smaller lattice (9 × 9) for up to 20 steps to ensure that the conceptualization of moral hypocrisy and the implementation of the strategies are correct.

[b]Several windows of summation for determining the quasi-stable state have been tested. It has been shown that the GT population is sufficient in order to detect quasi-stability for all populations, i.e., the population sizes do not change any more with exception of random fluctuations.

various initial conditions have been evaluated for each setting of population and temptation density: p_r and p_t were chosen from the interval [0.1, 0.9] in steps of 0.1, as the endpoints 0 and 1 lead to trivial solutions (see Section 2.2.3). This allows the calculation of the relative size of each population x_1, x_2, x_3, and x_4, for each setting of the four parameters p_r, p_t, d_p, d_t, whereas x_1 is the fraction of GR agents, x_2 is the fraction of GT agents, x_3 is the fraction of BR agents, and x_4 is the fraction of BT agents. For some investigations, we calculated the total count over all initializations of p_r and p_t for each of the four populations.

Furthermore, 11 social strategies (see Table 2.2) have been installed in the model. Those are either pure forms or combinations[c] of three basic strategies: to avoid agents tempted to seek agents with good reputation, or to disclose a hypocrite—latter means changing its behavior from GT to BT. The third basic strategy has been implemented in a local or a global form, whereas latter models the effect of mass media, i.e., everybody learns about moral hypocrisy of an agent. The "avoidance" strategy follows the intuition that people tend to avoid wrongdoers, the "seek" strategy follows the intuition that agents with good reputation are role models, and the "disclose" strategy follows the intuition that hypocrisy made public damages the reputation of the agent.

Table 2.2 Description of the social strategies (1–4: basic strategies, 5–11: combination of basic strategies) that intend to defeat moral hypocrisy

Number	Description of strategy
1	*Avoid agents that are tempted*: Every agent that has either a GT or BT neighbor moves to the closest free cell on the lattice without such neighbors
2	*Seek agents with good reputation*: Every agent that does not yet have either a GR or GT neighbor moves to the closest free cell on the lattice with at least one such neighbor

[c]In the model validation procedure, the effect of sequencing strategies has been checked for scenario B (see Section 2.2.3), where maximal changes in population distributions per strategy are visible. This revealed that the position of strategies 3 and 4 in the sequence do not significantly influence the population distributions (<1%), whereas there is a sequence effect for strategies 1 and 2. For the analysis, the sequence that benefits the GT population the least has been chosen (2 before 1).

3	*Disclose hypocrite (local version)*: Whenever the majority of agents in a two-degree Moore neighborhood (24 cells) of a specified GT agent is non-GT, this GT agent becomes a BT agent
4	*Disclose hypocrite (global version)*: After specifying a GT agent: whenever the majority of all other agents is non-GT, this GT agent becomes a BT agent
5	First strategy 2, then strategy 1
6	First strategy 1, then strategy 3
7	First strategy 1, then strategy 4
8	First strategy 2, then strategy 3
9	First strategy 2, then strategy 4
10	First strategy 3, then strategy 2, then strategy 1
11	First strategy 4, then strategy 2, then strategy 1

Note: For abbreviations, see Table 2.1.

These strategies do not, of course, replicate the complexity of behaviors toward wrongdoers and hypocrites in the real world. For example, the appeasing effect of public confessions of wrongdoing is not considered. Rather, they serve as an approximation for understanding the effect of various social strategies and combinations of strategies that intend to reduce moral hypocrisy. In particular, we can assess whether a strategy that seems to be preferred in modern societies—namely disclosing hypocrites by means of mass media, in particular if they are prominent figures like the former New York Governor Eliot Spitzer or the professional golfer Tiger Woods—is indeed as successful as one may assume. The strategies described in Table 2.2 have been implemented in the basic model so that each agent in each update cycle acts according to the strategy after payoff-comparison. The effect of each strategy is measured by the sum of the differences in the population distributions compared to the benchmark—the version without strategies—for all initial conditions with respect to reputation and temptation probabilities.

2.3 Paradigmatic Scenarios

During pre-tests, we have calculated the population distributions using three strategies (1, 2, and 3) for various pairs of d_a and d_t

(42 in total) and for all initial settings of p_r and p_t in steps of 0.05 including the trivial states (p_r, p_t = 0 and p_r, p_t = 1). This revealed, as expected, that in the trivial states where only one behavior type is present no changes occur, and that generic states with very low and very high p_r and p_t do not display interesting behavior. For example, for $p_t \approx 1$, we observe absolute dominance of moral hypocrisy. Using the results of the pre-tests, we established four paradigmatic scenarios for further analysis (see Fig. 2.1):

- **Scenario A—Pre-Modern:** This scenario is characterized by a low population (d_a = 0.1) and a low temptation density (d_a = 0.05). It follows the intuition that pre-modern societies consisted of small groups with high social control minimizing the number of available temptations. In the benchmark where no strategy is implemented in the model, GR dominates the overall count, whereas BT has—when the initial conditions are suitable—a fair chance to become a population of relevant size, too.
- **Scenario B—Modern Agricultural:** This scenario is characterized by a low population (d_a = 0.1) and high temptation density (d_t = 0.5). It follows the intuition that modern agriculture consists of large farms with low population density that have access to all means of modern societies in terms of mobility, communication, etc., that tend to increase the "temptation space." In this benchmark model setting, all behavior types have a chance to dominate, depending on the initial conditions.
- **Scenario C—Brave New World:** This scenario is characterized by a high population (d_a = 0.66) and low temptation density (d_t = 0.05). It implements the idea of a city-state like Singapore, for example, with a tight control regime with respect to temptations. In this benchmark model setting, GR dominates, although GT has a significant ratio in the population count.
- **Scenario D—Sin City:** This scenario is characterized by a high population (d_a = 0.66) and high temptation density (d_t = 0.5). It implements the idea of a densely populated city full of temptations. As expected, GT is by far the largest population in the benchmark setting.

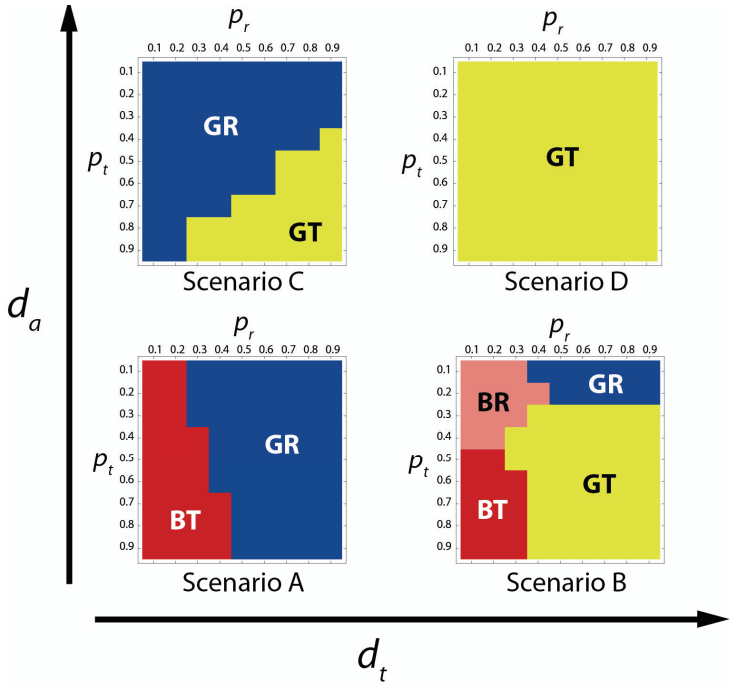

Figure 2.1 The four paradigmatic scenarios with respect to population and temptation density: Pre-modern (A), Modern Agricultural (B), Brave New World (C), and Sin City (D). For each scenario, the majority population (GR, GT, BR, and BT) after reaching quasi-stability is displayed in p_r/p_t-space according to the color code of Table 2.1. (The online version of this chapter has color graphics).

These scenarios are ideal types and only involve two of the many parameters that define human environments [9]. However, recent studies indicate that increasing population density appears to have catalysed the emergence of modern human behavior [14], making this parameter a natural choice for defining paradigmatic scenarios. And as temptation density—with respect to the number of available artifacts in different categories and behavior options, two main classes that define human environments [5]—is evidently an important factor with respect to the prevalence of moral hypocrisy, we suggest that these four ideal types of scenarios cover a wide range of possible human environments.

2.4 Defining a Measure for Population Diversity

We also analyzed the change in relative size distributions of the four populations x_1, x_2, x_3, and x_4 for specified parameter settings due to interventions by a measure for population uniformity $U(x_1, x_2, x_3, x_4)$ that is zero for $x_1 = x_2 = x_3 = x_4$ and maximal (i.e., 1) when only one population is present. The following measure fulfills these requirements:

$$U(x_1, x_2, x_3, x_4) = \frac{1}{6}\sum_{i,j=1}^{4}(x_i - x_j)^2 \tag{2.1}$$

When a strategy is compared to the benchmark with respect to population uniformity, we calculate:

$$\Delta U = \begin{cases} \left|U(\text{benchmark}) - U(\text{strategy})\right|, & \text{if no MC} \\ U(\text{benchmark}) + U(\text{strategy}), & \text{if MC} \end{cases} \tag{2.2}$$

MC stands for "majority change" that happens when for some setting of the model parameters, the population x_i is the largest in the benchmark, and $x_j \neq x_i$ is the largest after implementing the strategy. In this way, we compensate for the symmetry of U in x_i; otherwise, e.g., ΔU would be 0 when $x_j = 1$ in the benchmark and, after implementing a strategy, $x_j = 0$, whereas another population $x_j = 1$ completely dominates.

2.5 Results

We have calculated the population distributions for all four scenarios and for all 11 strategies (10 trials[d] per initial setting) both for the non-diversity and, in part, for the diversity model. In the following, we summarize the results of this analysis.

[d]Tests revealed that for 10 trials per setting of initial probabilities the standard deviation is usually in the order of less than 5% of the population size except for the case when a population has only very few members, where statistical fluctuations are naturally higher. As the populations usually are summed over all initial conditions, this effect is negligible.

2.5.1 Scenario Parameters Determine Population Distributions

The first insight provided by our model of moral hypocrisy is that the scenario parameters d_a and d_t determine the distributions of the four behavior types. In the non-diversity model, *no* strategy was able to change the majority of the dominating behavior type summed over all settings of p_r and p_t. In scenarios A and C, where d_t is low, the "good guys" (GR) dominate, whereas in scenarios B and D, where d_t is high, the "hypocrites" (GT) dominate. In scenario D, the dominance of GT is so overwhelming that no strategy was able to make any relevant impact. In the other scenarios, however, more pronounced effects are visible; exemplars of strategies that led to large changes in population distribution are displayed in Fig. 2.2. In particular, BT was able to build up a strong minority position both in scenarios A and B for selected strategies. In scenario C, GR was able to strongly increase its dominance for selected strategies.

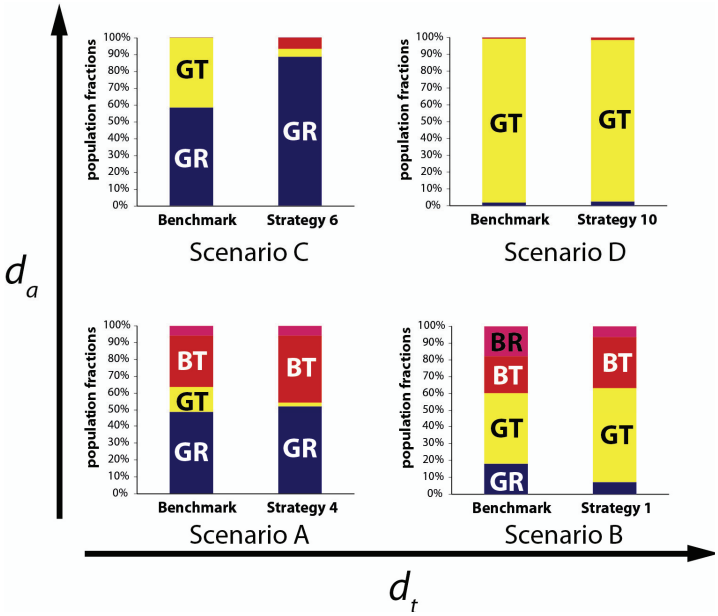

Figure 2.2 Effects of selected strategies on the population distribution in the four paradigmatic scenarios (examples): No strategy was able to change the dominance of the majority population.

2.5.2 Strategy Rankings and Conflicting Effects of Interventions

The effect of a strategy can be very different depending on the scenario in which it is implemented in the non-diversity condition. We show this by ranking the strategies according to their ability to increase the population of a specific behavior type relative to the benchmark population size (Fig. 2.3)—BR was excluded from this analysis, as this behavior type never was able to increase its population size in any strategy.

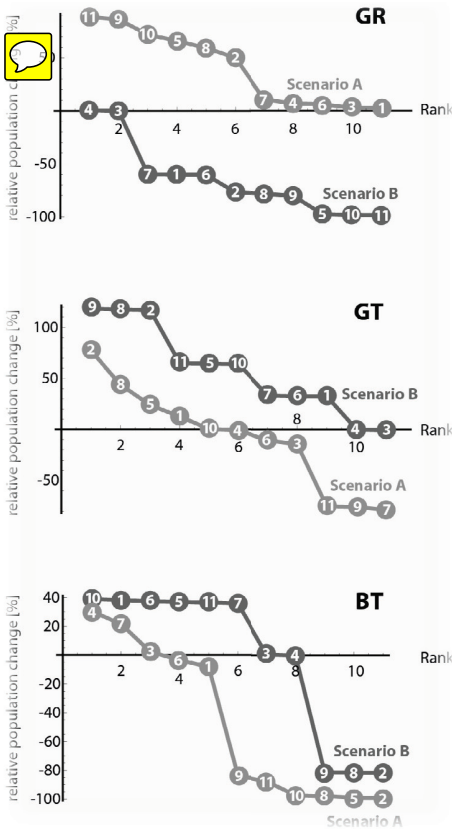

Figure 2.3 Ranking of strategies according to their effect on population size of GR, GT, and BT for scenarios A and B: Strategies optimal for scenario A are disastrous for scenario B. Numbers in dots indicate strategy according to Table 2.2.

For each ranking, we have calculated the Kendall Rank correlation—a measure of the similarity of rankings—and we display in Fig. 2.3. The two sequences with the highest dissimilarity for the "good guys" (GR), the "hypocrites" (GT) and the "bad guys" (BT). In all cases, the highest dissimilarity is obtained by comparing scenarios Pre-Modern (A) and Modern Agricultural (B). This reveals that strategies can have contradicting effects depending on the scenario. For example, strategy 11—a combination of disclosing hypocrites, seeking agents with good reputation and avoiding agents that are tempted—is optimal for GR in scenario A, but is disastrous in scenario B. The situation is analogous for strategy 9—a combination of seeking agents with good reputation and disclosing hypocrites—for GT: optimal in B, disastrous in A. For RT, strategy 10—again a combination of disclosing hypocrites, seeking agents with good reputation, and avoiding agents that are tempted—is optimal in B and disastrous in A. This means that the success of a strategy with respect to diminish the number of "hypocrites" strongly depends on the type of society, modeled by population and temptation density.

2.5.3 Strategy Effects Attributed to Four "Moral Worlds"

To get an overview of all strategies with respect to the intervention effect in the non-diversity condition, we grouped them in four classes based on how they change the population distribution: Good world strategies increase the GR population and decrease the BT and GT population, polarized world strategies increase the GR and the BT population, bad world strategies increase the GT and BT populations, and shiny good world strategies increase the GR and GT population (Fig. 2.4). We disregarded scenario D due to the small effects of all strategies in this case.

This analysis reveals tendencies with respect to the importance of main scenarios for strategy effects: First, the Modern Agricultural scenario B creates a context that promotes bad world strategies—i.e., BT and GT can often increase their weight. Second, the Brave New World scenario C creates a context that promotes polarizing world strategies increasing both the GR and the BT population. Third, the Pre-Modern scenario A creates a context that promotes shiny good world strategies.

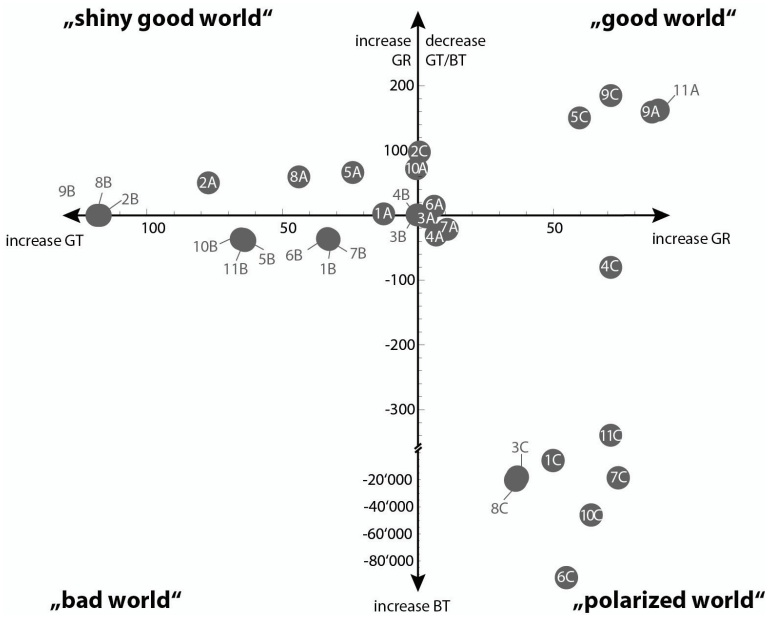

Figure 2.4 Classifying strategies in dependence of their effect in a scenario: The strong increase of the BT population in the polarized world for some strategies reflects the fact that this population is small in the benchmark. The axes display the relative population change in [%]. The axis separating "shiny good world" from "good world" has a different labeling for the right and left side.

With respect to the strategies themselves, no strategy can be attributed to only one strategy class. However, disclosing strategies in their pure form (3 and 4) tend to be polarizing, i.e., form strong minorities of BT agents. Strategy 2—seeking agents with good reputation—tends to be "shiny," which is plausible as the GTpopulation profits from a strategy that benefits good reputation.

2.5.4 Scenario-Diversity Overcomes Moral Hypocrisy

As d_a and d_t are the dominant factors with respect to population distributions in the quasi-stable state independent from the strategy, we analyzed the effect of strategies against moral hypocrisy on populations when the model implements *all four* basic scenarios simultaneously. This diversity condition can be

understood to increase the social complexity of the model, that is depending on the strategy agents may leave regions of high temptation density.

In the benchmark condition, GT ("hypocrites") is the dominant behavior for almost all initial settings, as expected by the payoff structure (Fig. 2.5). However, contrary to the findings in the non-diversity condition, 5 out of 11 strategies are able to change population majorities from GT to GR ("good guys"). Diversity seems therefore to be a crucial factor for overcoming moral hypocrisy.

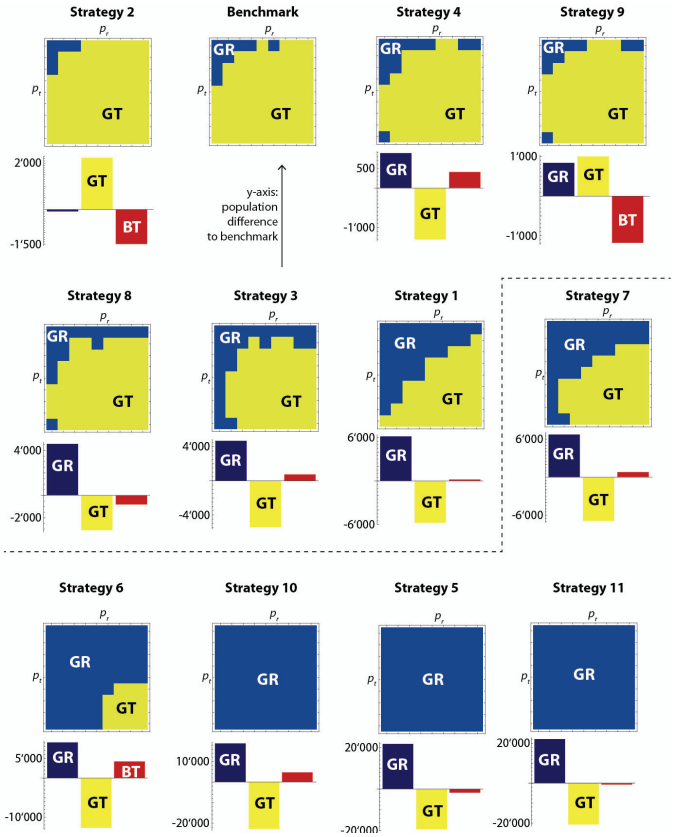

Figure 2.5 Effect of strategies in the diversity condition: Five out of 11 strategies, ranked from first row left to second row right according to increase in GR population, overcome the GT majority, indicated by dotted line. Shown are majorities in p_r/p_t-space and absolute changes in the population counts of GR, GT and BT.

It is remarkable that this majority change is caused mainly by the fact that agents move to regions where temptation density is low, although there is no built-in mechanism in the model to avoid temptations. Agents move away from other agents that are tempted, which is reflected by the fact that those strategies that include the basic strategy 1 ("avoid agents that are tempted") are successful with respect to majority change, although strategy 1 alone is not yet able to induce a majority change.

In order to further analyze majority changes in the diversity condition, we calculated the population uniformity p_r/p_t-space according to Eq. (2.1) and (2.2) for the benchmark and for all strategies. Figure 2.6 shows the mean $\Delta U(x_1, x_2, x_3, x_4)$ over all strategies compared to the benchmark. This reveals that the effect of strategies is not uniform in p_r/p_t-space. Rather, a low probability for having a good reputation p_r in the model initialization generally leads to larger population changes after strategy implementation, independent of the probability to be tempted p_t. Another local maxima in ΔU is discernible for high p_r and low p_t, i.e., an initial setting that consists mostly of GR agents ("good guys") is vulnerable for majority changes.

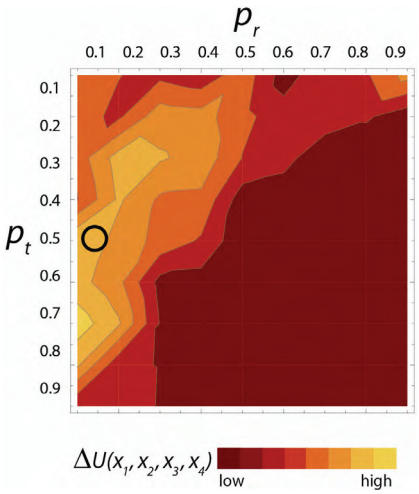

Figure 2.6 Mean changes in population uniformity over all strategies compared to the benchmark: The initial conditions affect the vulnerability of dominating populations for majority changes after strategy implementation differently. The circle indicates the parameter choice for the analysis depicted in Fig. 2.7.

Finally, we investigated the effect of population and temptation density for *fixed* p_r and p_t in the diversity condition. We changed both the total number of agents and temptations from 50 to 750 in steps of 50 by keeping the relative differences in densities in the four quadrants of the lattice. We fixed p_r and p_t such that maximal change in population uniformity can be expected—indicated by the circle in Fig. 2.6—and we chose strategy 5 as comparison to the benchmark, knowing that this strategy is able to change the benchmark majority (GT agents), see Fig. 2.5. In this way, we can analyze how population uniformity changes in d_a/d_t-space.

Figure 2.7 shows the result of this analysis. It reveals that for low d_a the "bad guys" BT are always the dominating population in the benchmark, whereas for a higher number of agents, d_t defines,

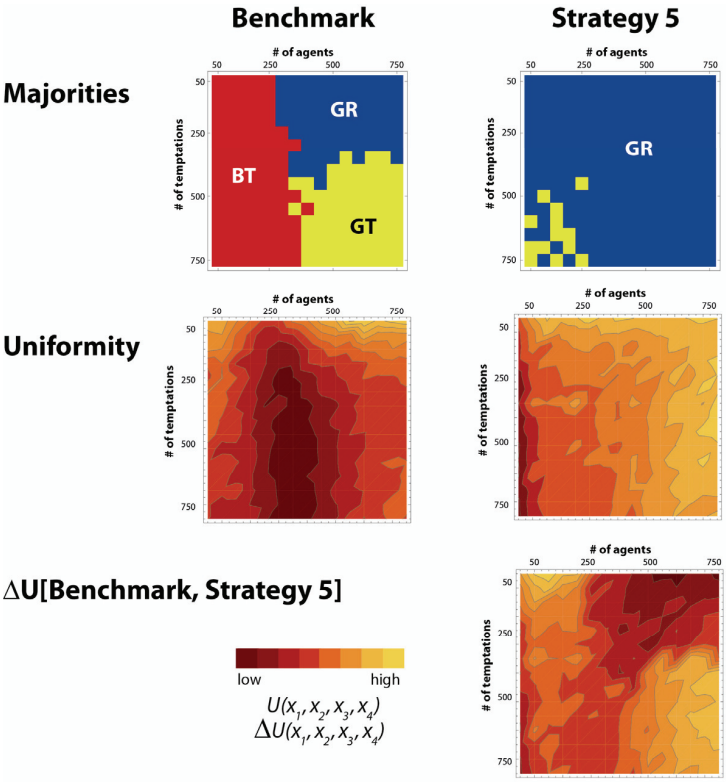

Figure 2.7 Effect of strategy 5 in the d_a/d_t-space: The change in population uniformity is highest for low agent and low temptation density as well as for high agent and high temptation density.

whether the "good guys" GR or the "hypocrites" GT dominate. Implementing strategy 5 changes, as expected, the majority count: GR becomes the dominating population. The zone with low population uniformity is shifted to the right. The highest changes in population uniformity are visible for the zone where both population and temptation density are either low or high. This means that strategy 5 is successful because it changes majorities in conditions that are usually not optimal for the "good guys" GR.

2.6 Discussion and Conclusion

This study shows that the success of strategies to overcome moral hypocrisy in a society strongly depends on the type of society in which the strategy is implemented. In the non-diversity condition, successful strategies in one scenario have disastrous consequences on the population of a specific behavior in other scenarios. Strategies thus have "side effects": For example, disclosing hypocrites in the Brave New World scenario may have the unwanted effect of also increasing the population of "bad guys." On the other hand, strategies against moral hypocrisy in a society with relatively low interaction density due to rather low population density, like the Modern Agricultural scenario, generally tend to be unsuccessful, indicating that not moral hypocrisy itself, but other aspects of social organization—e.g., interaction density—may be better targets for social interventions.

However, adding social complexity to the model changes the picture dramatically: If generic social scenarios are allowed to coexist, various strategies can overcome the dominance of moral hypocrisy. The key element for such a majority change is the strategy to "avoid wrongdoers" which has the indirect effect of moving the agents away from temptations, whereas the strategy "disclose hypocrites" does not have this effect. In this sense, our results confirm the importance of social complexity for the success of pro-social behaviors like cooperation [15,16]. Our results also accord with recent modeling studies that demonstrate the importance of agent mobility for enforcing pro-sociality [10]. It, furthermore, conforms to the philosophical insight that developing psychological competences or "virtues" to resist temptations in all circumstances may not be the best strategy to avoid unmoral behavior. Rather, a person should gain insights into the own moral psychology and—

based on this knowledge—avoid temptations for which the person is susceptible [6].

Certainly, this model will not allow finding straightforward applications for overcoming moral hypocrisy in real world societies. The primary reason for this is that the success of social strategies against moral hypocrisy depends on cultural and historical factors. Abundance of hypocrisy with respect to specific behaviors may, for example, have the effect that the moral norm itself is undermined and eventually vanishes—a phenomenon that is detectable for various norms [13]. We can also expect path-dependency with respect to the succession of several strategies against moral hypocrisy. For example, a society that first chooses to disclose hypocrites and later switches to an avoidance strategy may experience a different level of success in overcoming hypocrisy than a society that implements these strategies in reverse order. This model actually allows one to investigate path dependency of strategy implementations, for which preliminary results (not shown) indicate that this is another important factor in assessing strategies against moral hypocrisy.

We note several additional shortcomings of this approach that require further investigations: First, the current model does not take into account the inner psychological complexity of moral hypocrisy with respect to the type of temptations, for example, some agents may be tempted only by one kind of temptation. This simplification may also explain why the number of hypocrites even in a real world Sin City scenario is probably lower compared to the dominance of this population in the model. Although the model design allows in principle the integration of an inner agent psychology with respect to moral hypocrisy, it remains doubtful whether an extension of the model in this respect would indeed lead to additional insights, as no real world data is available to compare the model data of the prevalence of moral hypocrisy in a society in order to validate the refined model. Second, the model does not include the issue of "double standards" with respect to assessing the severity of moral hypocrisy, i.e., the fact that people tend to evaluate their own moral transgressions as less severe compared to the evaluations of the same transgressions committed by others. Including this aspect in the model would require a refinement with respect to agent psychology. Third, the current analysis does not involve all possible social

strategies against moral hypocrisy, i.e., it is incomplete in that respect. Fourth, there are other model parameters that may become the object of further investigations. Examples are changes in the payoff-structure or non-Moorean interactions between the agents. A non-exhaustive analysis of the effect for changes in the payoff-structure by weighting either moral reputation or temptation gain higher than the other component[e] reveals, however, no qualitative change of the results, as long as the major model assumption—moral hypocrites gain the most—is fulfilled.

In summary, the analysis provides indications that the abatement of moral hypocrisy cannot rely on simple and single strategies that abstract from the generic social setting in which hypocrisy emerges. Overcoming moral hypocrisy requires context-sensitivity in order to be successful and relies on societies that are both divers and allow for social mobility.

Acknowledgments

I thank Daniel Singer for his helpful advice in developing this model of moral hypocrisy. I also thank all members of the Institute for Advanced Topics in the Digital Humanities (UNC Charlotte) for their support.

References

1. Batson, C. D., Thompson, E. R., and Chen, H. (2002). Moral hypocrisy: addressing some alternatives. *J. Pers. Soc. Psychol.,* **83(2)**, 330–339.
2. Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., and Wilson, A. D. (1997). In a very different voice: unmasking moral hypocrisy. *J. Pers. Soc. Psychol.*, **72**, 1335–1348.
3. Benkler, Y. (2007): *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Yale University Press, New Haven).
4. Chriss, J. J. (2007). *Social Control: An introduction* (Polity Press, Cambridge).

[e]The payoff matrix has been changed as follows: emphasizing moral reputation: GT: +2, GR: +2, BT: +1; emphasizing temptation gain: GT: +2, GR: +1, BT: +2. In the benchmark condition, only the Brave New World scenario shows noticeable changes: when moral reputation is emphasized, the "good guys" profit even more, whereas when temptation gain is emphasized, the "hypocrites" gain more. This result is in accordance with the expectations.

5. Craik, K. H. (1971). The assessment of places. In *Advances in Psychological Assessment* (McReynolds, P. ed.), vol. 2, Science and Behavior Books, Palo Alto, CA, pp. 440–467.
6. Doris, J. M. (2002). *Lack of Character. Personality and Moral Behavior* (Cambridge University Press, Cambridge).
7. Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, **415**, 137–140.
8. Fointiat, V., Somat, A., and Grosbras, J.-M. (2011). Saying, but not doing: induced hypocrisy, trivialization, and misattribution. *Soc. Behav. Person.*, **39(4)**, 465–476.
9. Gärling, T. (1998). Introduction: Conceptualizations of human environments. *J. Environ. Psychol.*, **18**, 69–73.
10. Helbing, D., and Yu, W. (2009). The outbreak of cooperation among success-driven individuals under noisy conditions. *Proc. Natl. Acad. Sci. U. S. A.*, **106(10)**, 3680–3685.
11. Hrdy, S. B. (2009). *Mothers & Others: The Evolutionary Origins of Mutual Understanding* (Harvard University Press, Cambridge).
12. Marlowe, F. W., and Berbesque, J. C. (2008). More "altruistic" punishment in larger societies. *Proc. Biol. Sci.*, **275**, 587–590.
13. Nichols, S. (2004). *Sentimental Rules. On the Natural Foundations of Moral Judgment* (Oxford University Press, Oxford).
14. Powell, A., Shennan, S., Thomas, M. G. (2009). Late Pleistocene demography and the appearance of modern human behavior. *Science*, **324**, 1298–1301.
15. Santos, F. C., Santos, M. D., and Pacheco, J. M. (2008). Social diversity promotes the emergence of cooperation in public goods games. *Nature*, **454**, 213–216.
16. Shutters, S. T. (2011). Punishment leads to cooperative behavior in structured societies. *Evol. Comput.*, **20(2)**, 301–319.
17. Tong, E. M. W., and Yang, Z. (2011). Moral hypocrisy: of proud and grateful people. *Soc. Psychol. Person. Sci.*, **2(2)**, 159–165.
18. Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quart. Rev. Biol.*, **46**, 35–57.
19. Valdesolo, P., and DeSteno, D. (2007). Moral hypocrisy: social groups and the flexibility of virtue. *Psychol. Sci.*, **18**, 689–690.
20. Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.