How do Experiences Shape more and less Experienced Cops during Containment of Civil Violence?

Maikel Withagen (S1867733) [a] Steven Bosch (S1861948) [a]
Robin Kramer (S1970755) [a]

[a] *University of Groningen, Faculty of Mathematics and Natural Sciences, Nijenborgh 9, 9747 AG, Groningen*

1 Introduction

1.1 Problem

The civil war in Bosnia and Herzegovina from 1992 to 1995, the Ferguson unrest in 2014, and more recently the civil war in the Ukrain; Civil violence has been an issue in many countries for many years. There is no single cause for these riots and wars. Some were the result of cultural differences, some of the feeling of being treated injustly, and others came about due to political reasons. “Each war is as different as the society producing it" [8], and getting more insight into the development and handling of these riots is of utmost importance.

1.2 State of the Art

Much research has been done to understand how these riots emerge, including simulations of riots using a game theoretic approach [5] and social networks [4] among others. These investigations have shown that the behavior at a macroscopic level, that is, the behavior of an entire group, is the result of behavior on the microscopic level, that is, the individual agents of that group. In a more recent study, Goh and colleagues [3] also studied how macroscopic behavior emerged using a game theoretic approach in a simulation. The most important issue the authors adressed was how different events affected the individual tendency to riot and how this affected macroscopic behavior and situations.

Goh et al. included many different interactions in his simulation, such as a probability that civilians turned to active protesters, how jail time affected rehabilitation of arrested protesters, and how the amount and types of people affected individual decisions. The focus was thus mainly on the civilian and the protesters. What the authors failed to focus on, however, was how experience affected the cops’ behavior during riots. It is believed that many previous experiences ensure a more stable performance [1, 6], meaning that people switch less often between different strategies for the same goal. In other words, more experienced humans are less affected by new experiences compared to less experiened people, which can roughly be translated to the saying "you can’t learn old dogs new tricks."

1.3 New Idea

Knowing how experience influences group behavior may be of vital importance for the success of a group performance. For one individual, it is very well possible that the ‘correct’ course of

action may result in a negative outcome in a particular situation, because the world is not deterministic. The experienced person will be resilient to such an 'accidental' negative outcome, whereas a less experienced person may be affected more severely by that negative outcome. The latter person may therefore change to a less favourable course of action in a future occurence of said situation, which may result in frailty of the group's dynamics. It has already been shown in an organizational context that individual experience may determine the success of development of the organization [7]. Individual experience may thus have a great impact on group performance. As of yet this has not received much attention in a civil violence context however.

The current paper focusses on how experience influences cop behavior. By using a simulation of a multi-agent system, we were able to compare how experienced cop agents, not experienced cop agents and a mix thereof perform against a group of agressive hostile agents. A reinforcement learning strategy was used to let the individual agents learn the optimal strategy, which has been shown to be robust and natural way of teaching agents [2]. In the following sections, the simulation and data acquisition will be descibed in more detail. Following, the results of the different simulation will be presented. The paper concludes with a discussion of the implications of the results, and some shortcomings to this paper.

2 Method

2.1 Simulation model

Three types of agents will be included in this simulation: cops, hostiles and civilians. The goal of the cops is to keep (civil and cop) casualties as low as possible. The hostiles on the other hand have only one goal: to kill as many civilians as possible to cause mayhem and despair. Civilians have no particular function, but are subject to the actions taken by cops and hostiles. The simulation takes place in a 20x20 2D matrix. In every box of the matrix at most 50 agents can reside. The agents that reside in this box can only see the other agents in that box. This allows for easier computation of interactions with visible agents. In every simulation, around 15.000 agents ($\pm$ 5000 agents of each group) will be randomly spawned in the matrix.

2.1.1 Goals and Actions

The cops' goal is to keep casualties as low as possible. To achieve this, every cop must choose one of two actions. The cop can shoot a hostile, which eliminates a treath of future killings, but at the same time there is the risk of killing a civilian. The second action is to to save a civilian, who is then safe from being killed, with a higher risk of being killed by a hostile. Hostiles only have one action: to kill. Depending on the amount of civilians and cops, one group has a larger possibility to get killed by the hostiles. When, for example, only cops and hostiles reside in a box, the possibility that cops are targeted is 100 percent. It depends on the amount of hostiles how many civilians and cops may die.

		Shoot	Save
# Cops > # Hostiles	Many civilians	Ux	Uy
	Few civilians	Ur	Ut
# Cops < # Hostiles	Many civilians	Ua	Ub
	Few civilians	Uc	Ud

Table 1: The different scenarios an agent can encounter and fictional success values for each action.

All the cops, during every epoch, can either save or shoot once. Based on the overall result in that block, a reward is given to the individual agent. This reward can be either positive or negative. If many civilians die, for instance, a negative reward will be given, whilst when many hostiles were killed, a positive reward will be given. The decision that rewards are given based on group success, as opposed to individual success, is based on the idea that in a crowded situation it is not always clear what the results of individual actions are, but that a general idea of group performance can be perceived.

The reward is used to update the utility of the particular actions in a specific circumstance. These circumstances are shown in Table 1. The cops must thus decide, based on the amount of team members, opponents and civilians what action would be most successful, that is have the highest utility. In the beginning of the simulation, all the utilities are set to equal values, such that no biases may exist. Reward of an action can be calculated as following:

$$R(s,a) = \frac{Kills(s) + Saves(s') - Losses(s')}{Kills(s') + Saves(s') + Losses(s')}$$

in which R is the reward of action a in situation s, $Kills$ are the amount of killed hostiles in the new situation s', $Saves$ the amount of saved civilians, and $Losses$ the amount of killed cops and civilians. Based on these calculations the reward will lie between -1 (only deaths) and 1 (only kills and saves). Because hostiles only have one action, no learning is necessary.

As the Table 1 also shows, there is no situation in which the teams are of equal size. The cops must decide whether they have the overhand or not, which is done according to the following function:

$$\Omega = \frac{n_{cops} + (n_{cops} - n_{hostiles}) * \sigma}{n_{cops} + n_{hostiles}}$$

in which Ω is a value between 0 and 1, n is the amount of agents of a group in that box and σ is a random value between -1 and 1 to make the decision stochastic. It is assumed namely that the cops have no perfect knowledge of their environmnet. If Ω is higher than 0.5, the cops assume they have the overhand.

After each epoch, the agents can decide to move to a neighboring box in the matrix. If the agent decides to move it will go to the place, in which the action with the highest utility can be applied. In other words, the cops will move to where he/she will thrive best. If several

$$Will-This-Be-Done-With-Some-Function-With-Randomness-Question$$

2.1.2 Learning

Many reinforcement learning algorithms, such as Q-learning, have implemented the fact that newer experiences have a reduced influence [9]. This is implemented with a learning rate variable λ. High learning rates allow agents to learn more quickly compared to a low learning rate, but this also allows the agents to switch to suboptimal strategies more often. This may result in a lower performance compared to the low learning rate. Gradually changing the learning rate over time allows the function to converge to the true Q-value, that is the utility value. In the current simulation, a similar approach will be taken.

The utility will be updated according to the following function:

$$U(s,a)_{new}(t) = \frac{U(s,a)_{old} + \lambda * R(s,a)}{1 + \lambda}$$

in which $U_x(t)$ is the new utility (ranging from -1 to 1) of an action a in situation s, t is the moment in time, λ is the learning rate factor ranging from 0 to 1, and $R(t)$ is the reward, as

calculated before. After each epoch, in which one action could be taken by each agent, the utility functions will be updated.

To initialize the cop agents and set the learning rate variable, random values between 0 and 1 will be taken from a Gaussian distribution. This allows us to create a team with mixed levels of experience, which increases the realism of the simulation. By setting the skewness of the Gaussian distribution we can manipulate the mean learning rate value to be low and high, which allows the comparison of highly experienced teams (low mean learning rate) with less experienced teams (high mean learning rate).

2.2 Experiment Design

Four simulations were run. In the first simulation, learning rate values were drawn from a normal Gaussian distribution, having a mean value of 0.5. For two other simulations, the Gaussian distribution were postively and negative skewed, therefore moving the mean learning rate lower and higher respectivly. In the final simulation, the agents will start with a high learning rate value, which will be decreased over time. In other words, people gain experience on the job and will become less adaptive over time.

During the simulations, the following information was kept track of: (1) The number of killed hostiles per box, and the overall number of kills, (2) the number of overall dead cops and per box, and the amount of saved and killed civilians, (3) the number of saved civilians per box and the overall amount. The simulation of one riot is finished when all hostiles or cops have been killed, or when there are no civilians left in the area.

The simulations, as described above, were run ten times. One simulation is considered to be more successful if the mean successvalue (which is the same as the reward function) over these ten simulations was significantly higher compared to the other simulations.

3 Hypothesized Results

When comparing the first three simulation, in which the learning rate is fixed for the individual agents, we expect the following results to show. The group with a high learning rate (Group 1), will learning the successful strategies more quickly and adapt to that, compared to the two groups with a lower learning rate. However, because the high learning rate is also associated with more strategy changes due to 'accidental' outcomes, it is likely that a less optimal strategy will be taken more often than in the other two groups. This would result in a lower overall performance. Following this logic, the average learning rate group (Group 2) will be the next to achieve their best performance, which is a little higher compared to Group 1. The low learning rate group (Group 3) will be the last to achieve their best performance level. Because Group 3 does not change to a suboptimal strategy that quickly, it will most likely show the best performance.

The group with a decreasing learning rate over time (Group 4) is expected to have the best of both characteristics. Whereas the agents will initially learn the optimal strategy more quickly, over time the utility of the strategies will converge to an optimal level. In other words, over time the agents will stick to the strategies that have shown to have a higher utility in the beginning of the simulation. It is thus expected that Group 4 will show the best performance.

References

[1] John R Anderson and Richard King. *How can the human mind occur in the physical universe?* Oxford University Press, 2007.

[2] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI/IAAI*, pages 746–752, 1998.

[3] Chi Keong Goh, HY Quek, Kay Chen Tan, and Hussein A Abbass. Modeling civil violence: An evolutionary multi-agent, game theoretic approach. In *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, pages 1624–1631. IEEE, 2006.

[4] Timothy R Gulden. Spatial and temporal patterns in civil violence: Guatemala, 1977-1986. *Politics and the Life Sciences*, pages 26–36, 2002.

[5] Roger B Myerson. Game theory: analysis of conflict. *Harvard University*, 1991.

[6] Shelley Nason and John E Laird. Soar-rl: Integrating reinforcement learning with soar. *Cognitive Systems Research*, 6(1):51–59, 2005.

[7] Ray Reagans, Linda Argote, and Daria Brooks. Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together. *Management science*, 51(6):869–881, 2005.

[8] Nicholas Sambanis. Do ethnic and nonethnic civil wars have the same causes? a theoretical and empirical inquiry (part 1). *Journal of Conflict Resolution*, 45(3):259–282, 2001.

[9] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.