# Causal Analysis of the Emergent Behavior of a Hybrid Dynamical System

**Marcel Kvassay[1], Ladislav Hluchý[1], Peter Krammer[1], Bernhard Schneider[2]**

[1] Institute of Informatics, Slovak Academy of Sciences
Dubravska cesta 9, 845 07 Bratislava, Slovakia
e-mails: {marcel.kvassay, hluchy.ui, peter.krammer}@savba.sk

[2] EADS Deutschland GmbH
Landshuter Straße 26, 85716 Unterschleißheim, Germany
e-mail: bernhard.schneider@cassidian.com

*Abstract: This paper reviews selected concepts and principles of structural causal analysis and adapts them for exploratory analysis of a hybrid dynamical system whose continuous dynamics are described by ordinary nonlinear differential equations. The proposed method employs partial derivatives in order to calculate "causal partitions" of the system's state variables, which make it possible to quantify the extent to which various causes can be considered "responsible" for the emergent behavior of the simulated system. Causal partitions can be processed by machine learning techniques (e.g. clustering and classification), and so facilitate meaningful interpretations of the observed emergent behaviors. The method is applied to the simulated emotions of fear and anger in humans, in a hybrid agent-based model of human behavior in the context of EDA project EUSAS.*

*Keywords: Hybrid system; Causal analysis; Emergent behavior; Agent-based simulation; Human behavior modeling*

## 1 Introduction

This article is a report of work in progress extending our earlier paper [3]. From one point of view it can be considered a case study of one hybrid dynamical system. From a wider perspective, it is an attempt to introduce a new kind of analysis inspired by structural causality into the field of simulation studies. We demonstrate how structural causality facilitates meaningful interpretations of the emergent behaviors of complex systems and helps pinpoint their causes.

The paper is organized as follows: the rest of the Introduction provides a brief outline of structural causal analysis, while Section 2 adapts and applies its principles to the hybrid system under consideration. Section 3 details the simulation scenario selected for the experimental verification of the proposed approach. Section 4 describes the clustering and classification methods used to analyze the data and establish the relevance of causal partitions. Section 5 discusses these findings, proposes further improvements to our approach, and presents the first tentative results after the improvements were implemented.

## 1.1 An Outline of Structural Causal Analysis

Causality is one of the perennial topics in philosophy. Relatively recently, it has matured into a mathematical theory with significant applications in various fields of science. Although there are several competing accounts of causation, this paper focuses primarily on the comprehensive structural approach formulated by Judea Pearl and others [9, 1, 2], which subsumes and unifies the probabilistic, manipulative, counterfactual, and other specialized approaches. This section broadly follows the account given by Pearl in [9, 8, 11]. A detailed guide on how to perform structural causal analysis in practice is provided in [10].

According to Pearl[1] in his seminal book on causality [9], causal analysis can be applied to systems described by equations of the form

$$x_k = f_k\,(pa_k,\, u_k), \quad k = 1 \ldots n, \tag{1}$$

where $pa_k$ stands for the set of "parent variables" of $x_k$ directly determining its value through an autonomous mechanism captured by $f_k$, and $u_k$ stands for the effect of omitted factors. The autonomy of the mechanisms means any of them can be changed by external intervention without affecting the others. A set of such equations is called a "structural model." If, in addition, each variable (apart from the error terms $u_k$) appears on the left-hand side of some equation, then the model is called a "causal model." The error variables $u_k$ are also termed *exogenous* or *background*; they are simply considered as given. The variables $x_k$ are termed *endogenous*, i.e. determined by the equations within the system. A given value-assignment to the background variables constitutes a *world* or *context* in which the solution to the model equations is sought. In this paper, we restrict our attention to the recursive systems (systems without feedback loops), which possess a unique solution for each context. The equality sign in structural equations is endowed with directionality and is closer to the assignment operator in programming languages than to the standard algebraic symbol of equality.

Each causal model is associated with a causal diagram – a directed graph in which, for each equation, arrows point from $u_i$ and the parent variables in $pa_i$ toward their child, the dependent variable $x_i$. In fact, certain questions are more easily answered from the diagram than from the equations. In order to illustrate

[1] J. Pearl, *Causality: Models, Reasoning and Inference* 2nd Ed. 2009, p. 27 © Judea Pearl 2000, 2009, published by Cambridge University Press, reproduced with permission

this, here is an example with three pairs of equations reproduced from the Epilogue[2] to J. Pearl's book on causality [9]:

$$\begin{aligned} &Y = 2X \\ &Z = Y + 1 \end{aligned} \tag{2a}$$

$$\begin{aligned} &X = Y/2 \\ &Y = Z - 1 \end{aligned} \tag{2b}$$

$$\begin{aligned} &2X - 2Y + Z - 1 = 0 \\ &2X - 2Y - 3Z + 3 = 0 \end{aligned} \tag{2c}$$

These three pairs are algebraically equivalent, in the sense of having the same solutions, but only the first two pairs (2a) and (2b) qualify as structural models. While each equation in the third pair (2c) can be expressed as a linear combination of the equations in the preceding pairs, neither qualifies as structural, because it is not clear which variable is the dependent one (the child) and which are the independent ones (its parents). Moreover, even the first two pairs (2a) and (2b) do not describe the same causal model: their circuit representations using adders and multipliers shown in Fig. 1 make it obvious that the flow of causality in the second pair (Fig. 1b) is reversed with respect to the first (Fig. 1a). As such, we get different predictions from these two models concerning hypothetical interventions, e.g. "What happens if we set the value of the middle variable *Y* to *0*?"

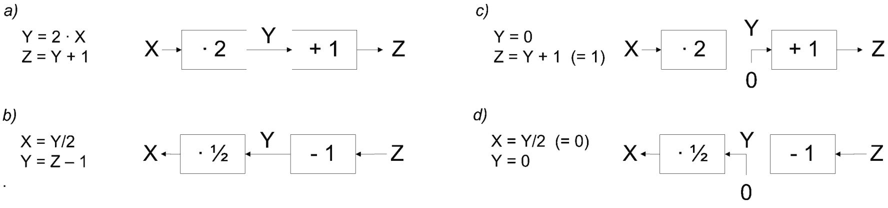

Figure 1

Circuit representations of equations (2a) and (2b) before and after an external intervention – setting Y to 0 (adapted from J. Pearl, *Causality: Models, Reasoning and Inference* 2nd Ed. 2009, pp. 416-7 © Judea Pearl 2000, 2009, published by Cambridge University Press, reproduced with permission)

In the first pair (2a), manipulating *Y* will affect *Z*, while in the second (2b), manipulating *Y* will affect *X* and leave *Z* unconstrained. This is shown both graphically and symbolically in Fig. 1c and 1d. Symbolically, the external intervention on *Y* means replacing the equation in which *Y* is the dependent variable with a new equation, in this case a straightforward value substitution *Y = 0*. The modified equations are then solved algebraically in order to determine the

[2] J. Pearl, *Causality: Models, Reasoning and Inference* 2nd Ed. 2009, pp. 416-7 © Judea Pearl 2000, 2009, published by Cambridge University Press, reproduced with permission

response of the model to the intervention. The first model (Fig. 1c) gives $Z = 1$, while the second gives $X = 0$ (Fig. 1d). In this way the structural approach provides a clear and unambiguous definition of causality. In Pearl's own words: "*Y* is a cause of *Z* if we can change *Z* by manipulating *Y*, namely, if after surgically removing the [original structural] equation for *Y*, the solution for *Z* will depend on the new value we substitute for *Y*." In contrast, the third pair of equations (2c) provides no causal structure or "intervention" guidance at all: it simply represents two general constraints on three variables, without telling us how the variables influence each other.

The directionality of causality has been one of the main obstacles to capturing causality satisfactorily in purely logical and purely statistical frameworks. The notion of cause can be further refined by distinguishing the necessary and sufficient aspects of causation, and the type-level from token-level causation. Another important concept linked with the token-level causation is that of an "actual cause" [1, 2]. For the purposes of this paper, however, these distinctions are not crucial. We shall therefore proceed with an example taken from [11], which demonstrates the use of structural models for various kinds of inference and for evaluating the effects of interventions.

The example, titled "The Impatient Firing Squad," analyses a fictitious scene just before the execution of a prisoner. The firing squad comprises a Captain and two riflemen. For the purposes of this analysis, the situation is modeled by 5 binary propositional variables: *U* ("Court orders the execution"), *C* ("Captain gives the signal"), *A* ("Rifleman-A shoots"), *B* ("Rifleman-B shoots"), and *D* ("The prisoner dies"). It is assumed that both the riflemen are law-abiding (i.e. they will only shoot if the Captain gives the signal) and competent (i.e. if any of them shoots, the prisoner will die). Likewise, the captain will signal only if ordered to do so by the court. These dependencies, expressed in the form of Boolean structural equations, lead to the following causal model:

$$C = U \tag{3a}$$

$$A = C \tag{3b}$$

$$B = C \tag{3c}$$

$$D = A \vee B \tag{3d}$$

In this model, *U* is *exogenous*, because it does not have its own structural equation in which it would appear on the left-hand side. Because this is a recursive system without feedback loops, the value of *U* uniquely determines the values of the remaining (*endogenous*) variables *C, A, B, D*. The causal diagram associated with this model is depicted in Fig. 2a.

It is obvious that in this model there exist just two consistent truth valuations of its variables: either they are all true, or they are all false, which greatly simplifies the task of evaluating the truth or falsity of the following "test" sentences:

S1: $A \Rightarrow D$ (“If rifleman-A shot, the prisoner is dead.”)

S2: $\neg D \Rightarrow \neg C$ (“If the prisoner is alive, then the Captain did not signal.”)

S3: $A \Rightarrow B$ (“If rifleman-A shot, then rifleman-B shot as well.”)

S4: $\neg C \Rightarrow D_A$ (“If the captain gave no signal and rifleman-A decides to shoot, the prisoner will die.”)

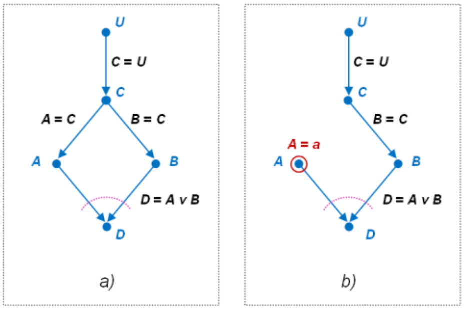

Figure 2

Causal diagrams for “The Impatient Firing Squad” example: a) original model, b) modified model after intervention (adapted from [11])

S1 is prospective inference from causes to effects (prediction), S2 retrospective or diagnostic inference from effects and back to causes (abduction), and S3 inference through common causes (transduction). Sentences S1 – S3 do not involve modifications of the original model and can be proved by purely logical means. Taking advantage of the fact that either all the variables are true, or they are all false, it is enough to check the truth values of S1 – S3 in these two settings. Since they all hold in both, we can conclude that they are entailed in the model.

S4 (“action”) is an instance of “true” causal thinking – it is like prediction except that it must be evaluated in a modified model. S4 states that rifleman-A shoots *without Captain’s signal*, which implies that the equation (3b) no longer holds. This is graphically depicted in Fig. 2b, where the link between $C$ (the Captain’s signal) and $A$ (the rifleman-A’s shot) is severed. Instead, the variable $A$ is assigned the value representing the rifleman-A’s new behavior: $A = True$. At the same time, S4 stipulates that the Captain did not signal, i.e. $C = False$ ($\neg C$ for short). Since S4 does not indicate any departure from the norm for the Captain, we assume that (3a) holds, i.e. $U$ is false. Likewise, by (3c), $B$ is false as well. However, equation (3d) still makes $D_A$ true because $A = True$ (the subscript “$_A$” in “$D_A$” denotes the value of $D$ in the minimally modified model, in which the equation for $A$ was replaced with substitution $A = True$). In this modified setting, S4 holds, because both its antecedent and its consequent are true. It needs to be kept in mind that, while evaluating S4, no other hypothetical modifications to the original causal model besides those explicitly stipulated by S4 are permitted.

This example also illustrates the distinction between *events* and *actions*. In the context of the model (3a) – (3d), if rifleman-A shoots in response to Captain's signal, it is an event, not an action. Events represent model variables assuming particular values from their allowed range. Thus, if the equations (3a) – (3d) yield the result "*A = True*", this would denote the event "Rifleman-A shoots," while "*A = False*" would denote the event "Rifleman-A does not shoot." The term "action" (or "intervention") is reserved for happenings that entail the modification of the model itself. For example, in the "action" sentence S4, the rifleman-A's decision to shoot *without Captain's signal* implies a disruption of one of the causal mechanisms – equation (3b) – and its replacement by another (in this case by a straightforward value substitution "*A = True*"). This is different from *A* assuming the value "*True*" under "normal" conditions described by (3a) – (3d).

These principles can be used to analyze dynamic situations as well. For dynamic analysis, however, the state variables describing the system need to be discretized with respect to time. Pearl in [9] and [11] provides an example of two forest fires advancing toward a house. In that example, the discretization is both temporal and spatial, since the changing state of the forest over time is conceptualized as a directed graph (causal diagram), in which each node represents the state of one patch of forest at a certain location *x* and time *t*. Pearl demonstrates a technique of "causal beams" through which it is possible to determine which of the two forest fires is the *actual cause* of the destruction of the house. This dynamic example hints at the potential value of structural causal analysis for simulation studies in general. In the model of human emotions that we analyze below, we apply these principles to ordinary nonlinear differential equations and the discrete dynamics that comprise our hybrid system.

# 2 Causal Analysis of a Hybrid System

The first point that needs to be addressed is whether our human behavior model (described in more detail in section 2.1) meets the criteria set for structural causal models. The main difference is that this model includes a particular kind of ordinary nonlinear differential equations with first derivatives with respect to time *t* on the left-hand side. In general, following the notation used in the definition (1), such equations can be written as

$$dy_k/dt = g_k\,(pa_k', u_k) \qquad k = 1\ldots n \tag{5}$$

where the function $g_k$ can be nonlinear. Another difference is that the dependent variable $y_k$ typically influences its own time derivative, i.e. it belongs to its own "parent set": $y_k \in pa_k'$. In order for such models to qualify as "causal", the differential equations need to be converted into difference equations, e.g. by replacing the derivatives with difference quotients:

$$(y_{k,j} - y_{k,j-1})/(t_j - t_{j-1}) = g_k\ (pa'_{k,j}, u_{k,j}) \qquad k = 1\ldots n, j = 1\ldots m \tag{6}$$

where $k$ indexes the original variables $y_k$ and $j$ indexes the discretized moments of time $t_j$, so that $y_{k,j}$ stands for the value of variable $y_k$ at time $t_j$. Denoting $t_j$ - $t_{j-1}$ as $\Delta t$, this can be rewritten as

$$y_{k,j} = y_{k,j-1} + \Delta t\ .\ g_k\ (pa'_{k,j}, u_{k,j}) \qquad k = 1\ldots n, j = 1\ldots m \tag{7}$$

In this form the relationship to (1) becomes clear. Variable $y_{k,\,j}$ at the left-hand side of (7) corresponds to the variable $x_k$ in (1), which means that in this “structural form” the value of a given variable $y_k$ at a given point of time $t_j$ is considered a separate “structural” variable, distinct from the value of the same variable $y_k$ at other points in time. Similarly, the whole right-hand side of (7) corresponds to $f_k\ (pa_k, u_k)$ of (1). Thus the relationship between the parent sets of (1) and (7) can be formulated as $pa_k = pa'_{k,\,j} \cup \{y_{k,\,j-1}\}$. Most importantly, in this form the parent set of the variable $y_{k,\,j}$ no longer contains this variable, but only the preceding values of $y_k$ in time, which are now considered different variables. Thus, after the discretization, no “structural” variable depends on itself. Therefore we can conclude that our human behavior model, after the discretization, does qualify as a structural causal model, as defined in [9]. Because we are interested in developing a general method for causal analysis of this kind of systems, we shall not go into details of our equations for *fear* and *anger*, but rather show a numerical solution for the general case. Yet, in order to understand our simulation experiments, an overview of the dynamics of our simulated agents will be helpful.

## 2.1 Internal Dynamics of Civilian Agents

A simplified diagram of our model is shown in Fig. 3. It represents the key factors affecting the behavior of civilian agents in the simulation scenario chosen for our case study. This model was used in project EUSAS financed by 20 nations under the *Joint Investment Program Force Protection* of the European Defence Agency. Interested readers can refer to a detailed exposition with a motivating example given in [5, 4]. Below we provide a brief summary of the model.

In line with the PECS modeling methodology [14, 13] used in project EUSAS, the agent behaviors are conceptualized as sequences of atomic, uninterruptible actions, e.g. one step in a certain direction, a single provocation or a single threat. Each behavior pattern is activated when its triggering motive becomes the strongest. For aggressive behaviors, the typical (but not the sole) triggering motive is *anger*, for fearful ones (such as withdrawal or flight to safety) the motive is *fear*. Regarding the dynamics of the simulated motives *fear* and *anger*, a convenient starting point is the top left corner of Fig. 3: the number of people surrounding the agent, their actions and other events in the vicinity affect the agent’s motives (*fear*, *anger*) as well as its other internal parameters (*arousal*, *readiness for aggression*). Besides events and actions, there is also a direct *social influence* of other agents on the agent’s *fear*, *anger* and *readiness for aggression*.

This is modeled according to Latané's formula of strength, physical proximity and the number of influencing agents [7].

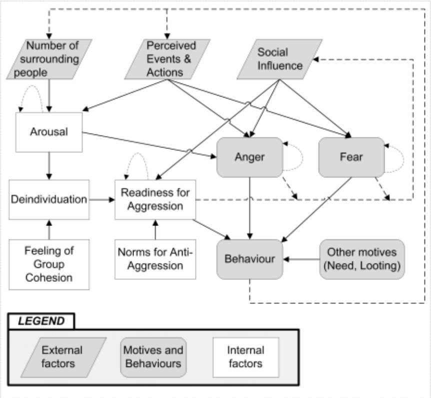

Figure 3
Sociopsychological model for the emergence of collective aggression in the form of a causal diagram

The internal *arousal* of the agent depends on the number of people in the vicinity and the violence of their actions. Speaking qualitatively, the higher the number and the more violent their actions, the sharper the increase of the agent's arousal. *De-individuation* means the agent considers himself a part of the crowd and no longer a separate individual, so the higher the agent's arousal and the more he feels a part of the group, the higher the de-individuation. *Readiness for aggression (RFA)* is jointly affected by the norms for anti-aggression, de-individuation and social influence as follows: (a) the higher the norms for anti-aggression, the lower the *RFA*; (b) the higher the de-individuation, the higher the *RFA*; and (c) the more social influence tends towards aggression, the higher the *RFA*. Psychological theories make aggression depend primarily on the *RFA*: without it, even a very angry person would not behave aggressively.

In [4] we tried to answer the question whether, in a scenario with this model, indirect social influence of nearby people was more important than external events in shaping the civilian agents' behavior. The answer provided by the causal summary of the scenario indicated that external events were more important. In this paper we revisit the same scenario (see section 3 for details) with another, more difficult problem: while experimenting, we have noticed that for a particular parameter setting, the emergent collective behavior of the civilians developed along two sharply diverging trajectories. In some cases, almost all of them got afraid and left the scene, while in others, almost all got angry and joined the attack on the security forces. Because our model incorporates an element of randomness, a certain variation of the emergent behavior of the civilians was to be expected, but the extreme variation that we witnessed was unusual and called for an explanation. Since the leading question was "Why is this happening?" it provided a welcome opportunity to test the relevance and practical utility of our approach.

## 2.2 Causal Partitioning – A Brief Outline

Following [4], we shall briefly illustrate causal partitioning on motive *fear*. Its continuous dynamics is driven by the equation

$$dF/dt = f(F(t), I_F(t)) \tag{8}$$

where $F$ stands for Fear, $I_F$ for fear-related social influence, and $f$ for a nonlinear function of these two arguments. In Euler numerical method this leads to

$$F(t + \Delta t) \approx F(t) + \Delta t \,.\, f(F(t), I_F(t)) \tag{9}$$

where $\Delta t$ is the simulation time step. $F(t + \Delta t)$ stands for the new "continuous" value of Fear, to which the discrete part of the dynamics ($E_F$) is yet to be added:

$$F_T(t + \Delta t) = F(t + \Delta t) + \Delta E_F \tag{10}$$

where $F_T$ stands for the new total value of Fear and $\Delta E_F$ for the cumulative fear-related impact of the external events perceived by the agent during the time interval ($t$, $t + \Delta t$>. This means we model the external events as taking the effect at the end of the time interval during which they occur.

Causal partitioning starts by linearizing the function $f$ through its partial derivatives with respect to its two parameters $F$ and $I_F$:

$$f_j \approx f_{j-1} + \partial f/\partial F \,.\, \Delta F + \partial f/\partial I_F \,.\, \Delta I_F \tag{11}$$

where $\Delta F$ and $\Delta I_F$ stand for the differences in the values of $F$ and $I_F$, respectively, between the start and the end of the time interval ($t - \Delta t$, $t$>. For brevity we have switched to indexing, where $f_j$ represents the current value $f(t)$ and $f_{j-1}$ the previous one, $f(t - \Delta t)$. Next, we exploit the fact that $\Delta F$ over the time interval ($t - \Delta t$, $t$> equals the sum of the discrete and the continuous changes during that period, i.e. $\Delta F = \Delta E_F + f_{j-1} \,.\, \Delta t$. Substituting this into (11) yields

$$f_j \approx f_{j-1} + \partial f/\partial F \,.\, (\Delta E_F + f_{j-1} \,.\, \Delta t) + \partial f/\partial I_F \,.\, \Delta I_F \tag{12}$$

which can be rewritten as

$$f_j \approx C_1 \,.\, f_{j-1} + C_2 \,.\, \Delta E_F + C_3 \,.\, \Delta I_F \tag{13}$$

with $C_1$, $C_2$ and $C_3$ representing the weighting factors that equal, respectively, $1{+}\Delta t.\partial f/\partial F$, $\partial f/\partial F$, and $\partial f/\partial I_F$. These factors can be evaluated numerically. As we explained in [4], this formula is the basis for our algorithm for causal partitioning of *fear* into causal partition vectors, where each vector component represents the quantitative contribution of one "causing" factor. The first derivative of fear $f$ at any moment is then represented by the causal partition vector ($f_E$, $f_I$), whose components sum up to $f$. The first component $f_E$ stands for the cumulative contribution of the external events $E_F$, the second $f_I$ for the cumulative contribution of fear-related social influence $I_F$. Analogously, the value of fear $F$ can be represented by the partition vector ($F_E$, $F_I$) whose components sum up to $F$. The causal partitioning algorithm proceeds roughly as follows:

To get the current partition of the first time derivative of fear ($f_j$), take its previous partition $f_{j-1} = (f_{E,\,j-1}, f_{I,\,j-1})$, multiply its members by $C_1$, then add the contributions of the external events and social influence over the interval ($t$ - Δ$t$, $t$> as per (13) to their respective partition components $f_E$ and $f_I$.

Next, to get the new partition for Fear ($F_{j+1}$), take its current partition ($F_j$) and add to it, component by component, the increment as per equation (9) using the current partition of its first time derivative $f_j$. The increment is a vector ($\Delta t \,.\, f_{E,\,j}$, $\Delta t \,.\, f_{I,\,j}$).

Last, add the cumulative value of the external events perceived during the new step ($\Delta E_F$ for time interval ($t$, $t + \Delta t$>) directly to the $F_E$ component.

In this simplified account we have assumed the zero initial value of Fear ($F_0$). If it is non-zero, it qualifies as a separate causal factor and requires a dedicated component in the causal partition vector ($F_{F0}$). Thus, the value of Fear is in fact partitioned into a casual partition vector ($F_{F0}$, $F_E$, $F_I$).

The analysis of Anger proceeds analogously. The continuous part of its dynamics is driven by the equation

$$dA/dt = g(A(t), L(t), I_A(t)) \tag{14}$$

where $A$ stands for Anger, $L$ for Arousal, $I_A$ for anger-related social influence, and $g$ for a nonlinear function of these arguments. The Euler method then leads to

$$A(t + \Delta t) \approx A(t) + g(A(t), L(t), I_A(t)) \,.\, \Delta t \tag{15}$$

Again, to this new “continuous” value of Anger, the discrete anger-related impacts of the events perceived during the time interval ($t$, $t + \Delta t$> have to be added:

$$A_T(t + \Delta t) = A(t + \Delta t) + \Delta E_A \tag{16}$$

A process of causal partitioning analogous to that for Fear then partitions $A_T$, the new total value of Anger, into a casual partition vector ($A_{A0}$, $A_E$, $A_I$, $A_L$)

In general, a causal partition of a model variable $X$ is a vector-like structure ($X_1$, $X_2$ ... $X_n$) whose components sum up to $X$ and where each component represents the portion of the value of $X$ attributed to one specific factor. This makes it possible to quantify the extent to which various factors can be considered “responsible” for the value of $X$ at any given moment in a given simulation scenario.

Concerning the simulation experiments in this paper, by a causal summary of a simulation run we mean the causal partition vectors representing the final values of *fear* and *anger* averaged over all the civilian agents in the scenario. This data, supplemented with a few other attributes, is then passed on to the machine learning algorithms for further analysis.

# 3 Simulation Experiment Scenario

As mentioned, we revisit the scenario from [4], in which a crowd of civilians is looting a shop, and the approaching soldier patrol is supposed to stop the looting and disperse the crowd. The scene is shown in Fig. 4. Black areas represent buildings and barriers unreachable to agents. The rectangle with gray interior near the top is the shop being looted. It is surrounded by dots, each representing one agent. The dark ones are the looters; the light-colored ones are the violence-prone individuals, whose intention is to attack the soldiers. The soldiers are represented by the three medium gray dots in the bottom part of the figure.

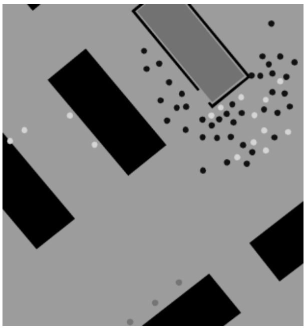

Figure 4

Initial stage of the simulation scenario

Civilian agents are endowed with one “default” motive and a matching behavior by which they try to satisfy it. For looters this leads to “looting” and for the violence-prone individuals to stone-pelting the soldiers. Additionally, the agents monitor what happens around them, which may excite fear or anger, in which case they start behaving fearfully (i.e. run away) or aggressively. As the patrol nears, this may induce fear in some looters who then start leaving the scene. The violence-prone individuals, however, do not get afraid but attack the patrol. The violence may impact the remaining looters in two possible ways – they may either get afraid and leave, or get angry and join the attack. How many get afraid and how many get angry depends on various parameter settings. The key ones are the initial values of *fear* and *anger* (which were set to $F_0 = 0.3$ and $A_0 = 0.2$) along with the main event impacts summarized in Table 1.

Table 1

Main event impacts on *fear* and *anger*

| Impacts: | Impact on Fear | | Impact on Anger | |
|---|---|---|---|---|
| Events: | Direct | Indirect | Direct | Indirect |
| Effective shot | 0.4 | 0.35 | 0.1 | 0.25 |
| Warning shot | 0.3 | 0.3 | 0.1 | 0.1 |
| Stone thrown | 0.002 | 0.002 | 0.18 | 0.15 |

The values of *fear* and *anger* are incremented by the event impacts shown in Table 1 every time the corresponding event is perceived by the agent. The "direct" values from Table 1 are used when the perceiving agent is the target of the event; otherwise the "indirect" values are used. What may seem counterintuitive at first, is that each event affects both *fear* and *anger*, but in different proportions. All the agent motives, including *fear* and *anger*, are real-valued and restricted to the closed interval <0, 1>. *Looting* motive is set to a constant value of *0.7*, so *fear* or *anger* must cross this level in order to affect the agent behavior. Sensory perception of the agents is limited both spatially (by a radius of 50 m for throwing stones and 150 m for gun shots) and emotionally (if the average of *fear* and *anger* crosses the level of 0.5, further sensory perception of events is blocked).

Unlike our civilians, our soldier agents are much simpler: they are just passing by and act in self-defense. Their rule of self-defense says that when a given civilian first throws a stone at a soldier, that soldier responds by a warning shot in the air. If the same civilian throws a stone at the same soldier a second time, the soldier is permitted to use an effective shot aimed at the legs of the attacker in order to immobilize him. That is, of course, an extreme simplification, but it proved useful in the early phases of project EUSAS for calibrating the civilian agents.

As part of the present case study, we ran this simulation scenario 300 times with the simulation time step of 300 milliseconds, and then again 300 times with the time step of 100 ms. This enabled us to gauge the effect of the time step size (and of the resulting discretization errors) on the observed emerging behavior of the agents. Since the time-evolution of our scenario is rather fast, it was sufficient for each simulation to cover just 90 seconds of simulated time. At the end of this period, the average values of *fear* and *anger* were recorded and their causal partition vectors (along with other relevant data) were passed on to machine learning algorithms for further analysis.

# 4 Machine Learning

Regarding data structure and pre-processing, our data consisted of two data sets, one for 100 ms and another for 300 ms time step. Each set contained 300 records with 12 numerical attributes: 7 components of causal partitions of average final values of *anger* ($A_{A0}$, $A_E$, $A_I$, $A_L$) and *fear* ($F_{F0}$, $F_E$, $F_I$), followed by 3 measures of effectiveness (MoE) used to evaluate scenarios in project EUSAS: $N_E$ (total number of effective shots), $N_W$ (total number of warning shots), and $N_S$ (total number of thrown stones). The last two attributes, *A-count* and *F-count*, stand for the cumulative numbers of times that *anger* and *fear*, respectively, became the strongest motives in some civilian agent. For scenarios that turn aggressive we expect a high *A-count* as well as high MoE values and a low *F-count*, while for the "timid" scenarios we expect a high *F-count* and a low *A-count* as well as low MoE values. We have included MoE as a sort of "competition" to our causal summaries: it is evident that MoE can classify the scenarios well, since aggressive developments imply high numbers of thrown stones as well as gunshots. MoE, however, lack the explanatory power: they do not tell us anything about why a particular scenario took an aggressive or a timid turn. We have verified our expectations by the clustering exercise shown in Fig. 5.

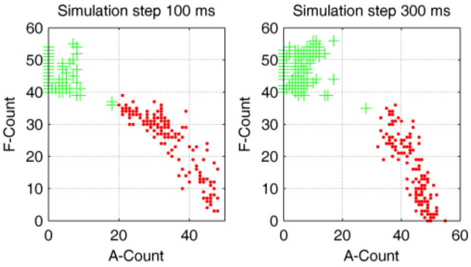

Figure 5

Data clustering into two clusters for 100 ms and 300 ms data sets

The charts have *A-count* on the x-axis, *F-count* on the y-axis and represent each simulation run as one data point. In order to improve readability, we have used the jitter method (adding noise with small amplitude). The left chart represents the 100 ms data set, the right one the 300 ms data set. As expected, in both data sets there seem to be two distinct clusters: the compact one in the top left corner ("timid" scenarios), and the elongated one in the bottom right part ("aggressive" scenarios). The elongated shape of the "aggressive" cluster is due to the varying levels of fearful behaviors that appeared alongside aggression. The elongation points towards the "timid" cluster because an increase in fearful behaviors goes hand in hand with a reduction in the aggressive ones. In other words, although our agents can switch between *fear* and *anger* several times during one scenario, this is rare and most of them switch just once from their "standard" motive to either *fear* or *anger*. This explains why simulations tend to cluster near the diagonal.

Both the clustering exercise and the classification experiments described below were executed in Weka [15], ver. 3.7.9, which uses expectation maximization algorithm. Clustering added a new attribute (cluster ID) to our data. On the scaled data thus pre-processed, we then trained several classification models, choosing the "cluster ID" as the target. Our classifiers were based on SVM (Support Vector Machine) with SMO algorithm (Sequential Minimal Optimization algorithm) [12]. Compared to other techniques, SVM is known to be more intuitively interpretable. For two clusters, an SVM model consists of a function describing the boundary hyperplane separating the clusters in the feature space. If the boundary function gives a positive value for some data point, then it belongs to one cluster, and if the value is negative, then it belongs to the other cluster. Zero value means the point lies on the boundary. In the exercise, we used the number of correctly classified instances as a quality measure in 10-Fold Cross Validation.

We start our classification along the lines of forward parameter selection in order to find out which of the causal attributes is the best standalone predictor of "cluster ID". For each attribute $X$, SVM produces a classification model of the form $aX + b$. The value of $X$ for which $aX + b = 0$ is the *cut-off value* defining the boundary: $X = -b/a$ (for $a \neq 0$). Models with $a = 0$ are trivial as they assign all the simulations to one cluster. The results of this exercise are shown in Table 2 (mark "--" in the place of a cut-off value means the model was trivial, having $a = 0$). Since all the attributes were normalized, all the cut-off values (so long as they exist) fall within the closed interval <0, 1>. For comparison, we have also included the "competing" SVM models based on MoE ($N_E$, $N_W$, $N_S$).

Table 2
Forward selection of the most important attributes on the basis of SVM models

| Attribute:<br>Measure: | $A_{A0}$ | $A_E$ | $A_I$ | $A_L$ | $F_{F0}$ | $F_E$ | $F_I$ | $N_E$ | $N_W$ | $N_S$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Cut-off value 100 ms | 0.45 | 0.24 | 0.48 | -- | 0.52 | 0.39 | 0.47 | 0.22 | 0.37 | 0.12 |
| Accuracy [%] 100 ms | 88.7 | 61.0 | 93.0 | 51.7 | 99.0 | 97.7 | 80.3 | 97.0 | 95.3 | 88.7 |
| Cut-off value 300 ms | 0.70 | 0.78 | 0.01 | 0.44 | 0.59 | 0.40 | 0.39 | 0.31 | 0.35 | 0.20 |
| Accuracy [%] 300 ms | 80.0 | 67.7 | 48.7 | 60.3 | 97.0 | 93.0 | 91.0 | 98.0 | 93.0 | 92.0 |

The big surprise for us was the high prediction accuracy of the causal partition component corresponding to the initial value of fear ($F_{F0}$): 99 % for 100ms dataset and 97 % for the 300ms dataset. If we were just looking for a high-quality classifier, we could have proclaimed our task finished at this point. For us, however, these classifiers are simply a source of hints about the underlying mechanism responsible for the observed divergence of the simulation trajectories (timid versus aggressive). And in this respect, as it turned out, things were much more complicated. In a straightforward interpretation the attribute $F_{F0}$ being the best predictor indicates that the initial setting of the value of fear ($F_0$) might be the underlying cause of the observed divergence of simulation trajectories. But we know this cannot be, since in all the 600 simulation runs this initial setting was kept the same ($F_0 = 0.3$). Thus, the fact that $F_{F0}$ does not remain constant must be due to other factors, most likely the other components of the causal partition of fear, to which it is tied by the partition definition constraint $F = F_{F0} + F_E + F_I$.

This brings us to the question of mutual correlations among the attributes, which we show in Table 3. Its lower triangular portion shows them for the 100 ms data set, the upper triangular one for the 300 ms data set. We see that $F_{F0}$ is indeed highly correlated with $F_E$ and $F_I$ in both data sets. We also note another redundant attribute $A_{A0}$ (highly correlated with $F_E$), which again cannot lead us towards the cause of the divergence, since the initial value of anger $A_0$ was the same ($A_0 = 0.2$) in all the simulations. Overall, this points towards $F_E$ as the primary factor, since it is also the second best standalone causal predictor of "cluster ID", with 97.7% accuracy for the 100 ms dataset and 93% for the 300 ms one. Before delving into possible explanations, however, we shall first try to confirm this finding by a process similar in spirit to backward parameter selection.

Table 3
Mutual correlations of attributes (lower-triangular section: 100 ms, upper triangular: 300 ms)

| | $A_{A0}$ | $A_E$ | $A_I$ | $A_L$ | $F_{F0}$ | $F_E$ | $F_I$ |
|---|---|---|---|---|---|---|---|
| $A_{A0}$ | 1 | 0.46 | -0.34 | 0.22 | 0.49 | -0.55 | -0.32 |
| $A_E$ | 0.54 | 1 | -0.50 | -0.28 | 0.44 | -0.40 | -0.41 |
| $A_I$ | 0.44 | 0.13 | 1 | 0.19 | 0.31 | -0.23 | -0.28 |
| $A_L$ | -0.17 | -0.67 | -0.14 | 1 | 0.44 | -0.47 | -0.41 |
| $F_{F0}$ | 0.72 | 0.36 | 0.76 | 0.11 | 1 | -0.89 | -0.88 |
| $F_E$ | -0.73 | -0.32 | -0.68 | -0.14 | -0.93 | 1 | 0.75 |
| $F_I$ | -0.35 | -0.33 | -0.53 | -0.12 | -0.71 | 0.52 | 1 |

The initial steps of backward parameter selection are shown in Table 4. We start with "complete" models with 7 causal attributes (top two lines) which reach 99% accuracy for both data sets. For comparison, we include the SVM models based on MoE in the next two lines. These perform slightly better for the 300 ms data set and slightly worse for the 100 ms one, thus confirming the quality of our causal models. We now remove the two "redundant" causal attributes $F_{F0}$ and $A_{A0}$, expecting that this should not impair the accuracy of our models, which is borne out by the last two lines in the table. Having got rid of high correlations, we can gauge the key factors among the remaining variables. In SVM models, these tend to be the ones whose coefficients in the boundary plane function have the highest absolute value. On this criterion, $F_E$ comes out as the most important attribute for both data sets – in total agreement with the first step of forward parameter selection in Table 2. In the 100 ms data set, the second and the third place belong to $A_I$ and $F_I$, respectively; while for the 300 ms data set their ranking is reversed. Although these additional attributes only add a few percent to the accuracy (since $F_E$ alone reaches 97.7% accuracy), we decided to explore in more detail the pair-wise combinations for additional clues they might offer.

Table 4
Trained SVM models with causal attributes and MoE

| Time Step | Function of the boundary plane | Correctly Classified |
|---|---|---|
| 100 ms | $- 0.10 * A_L + 0.23 * A_E + 1.44 * A_{A0} + 2.39 * A_I - 2.10 * F_E + 3.21 * F_{F0} - 0.79 * F_I - 2.22$ | 99.0% |
| 300 ms | $- 0.77 * A_L + 0.24 * A_E - 0.69 * A_{A0} + 2.04 * A_I - 1.65 * F_E + 4.51 * F_{F0} - 1.92 * F_I - 0.78$ | 99.0% |
| 100 ms | $4.53 * N_E + 3.41 * N_W + 1.98 * N_S - 2.36$ | 97.3% |
| 300 ms | $3.23 * N_E + 3.05 * N_W + 1.49 * N_S - 2.46$ | 99.3% |
| 100 ms | $0.47 * A_L + 1.03 * A_E + 3.16 * A_I - 5.16 * F_E - 1.61 * F_I + 0.77$ | 99.3% |
| 300 ms | $-0.43 * A_L + 0.85 * A_E + 2.96 * A_I - 4.49 * F_E - 4.16 * F_I + 2.74$ | 98.3% |

Considering the five causal attributes $A_E$, $A_I$, $A_L$, $F_E$ and $F_I$, there are ten possible pairs. In the 100 ms data set only one of them ($F_E$ *x* $A_I$) was found to outperform $F_E$ on its own. At the same time, this pair reached the same 99% accuracy as the full SVM model with seven causal attributes in Table 4. Thus, we can conclude that for the 100 ms dataset $A_I$ and $F_E$ together contain all the information present in the causal partitions regarding the aggressive versus the timid turn of our simulations. In the 300 ms dataset, the best pair is $F_I$ and $F_E$ reaching 97% accuracy, which comes close to the 99% accuracy of the full causal model. We discuss the implications of these results in the next section.

# 5 Discussion

Some of our causal SVM models reached very high quality, correctly classifying 99% of simulations, which means the causal partitions are good predictors of their "aggressive" or "timid" turn. On this basis, we feel justified in affirming the *relevance* of causal partitions for human behavior model exploration. As for their *practical utility*, this is more challenging, because by *practical utility* we mean their ability to guide us toward the aspect of the model which, if modified, would bring about the disappearance of the diverging trajectories – "aggressive" versus "timid" – for one input parameter setting. We do not expect our method to directly "compute" the answer, but rather assist us in the process of formulating and testing hypotheses. Toward this end, we first need to identify and interpret the key factors behind the good performance of our "causal" classifiers. Exploration along the lines of forward and backward parameter selection pinpointed $F_E$ as the primary candidate for explanation. Additionally, the interpretation of the reduced models in Table 4 supplied the second and third most important factors: $A_I$ and $F_I$. In general, we can therefore conclude that the most important factors influencing the trajectory of the simulations seem to be, first, external events acting through fear ($F_E$), followed by social influence acting through both anger ($A_I$) and fear ($F_I$). This is the kind of hint that machine learning techniques could extract from our causal partitions. In order to proceed further, we needed to incorporate deeper technical knowledge of our simulation model in our hypotheses.

Our initial hypothesis was that early in the scenario – as a result of some unknown process or a random fluctuation – there forms a nucleus of agents that are either angry or afraid (while the other agents are still under the influence of their standard motives) and this nucleus then "converts" the rest of the agents by their social influence. If this were the case, we would expect the social influence components $A_I$ and $F_I$ to be negatively correlated (i.e. working against each other) and at the same time to be the best predictors for classification. However, Table 3 shows only a small (albeit negative) correlation and, moreover, they are not the best predictors, since their combined accuracy was only about 95%. At this point of time our method was not yet so mature as to enable us to reject this hypothesis outright, but its likelihood certainly decreased. The main weakness of our approach was that we causally partitioned only the final values of fear and anger, while the really "decisive" period seemed to be the early part of the scenario. We needed to dynamically identify the moment in which the scenario divergence began and then apply the causal partition process at that point.

Our second hypothesis dealt with $F_E$ and the early attack by the violence-prone individuals. There was an element of uncertainty as to how many stones they would be able to throw. They select the closest soldier as their target, and if they hit him twice, they are immobilized by an effective shot. Thus, in the worst case, they only throw two stones, while in the best case, four (with three soldiers, the fourth stone throw always results in immobilization). Given that stone throws incite more anger, while the effective shots more fear, the proportion of the stone throws versus effective shots in the early part of the scenario might be the tipping factor determining its subsequent aggressive or timid turn. If this hypothesis were true, then by adjusting the soldiers to use only warning shots we should force all the scenarios to take the aggressive turn. We tested this experimentally, permitting soldiers only to use warning shots, but the two divergent trajectories still persisted. Thus the second hypothesis had to be discarded as well.

The above experiment also rendered unlikely our third hypothesis – that our agent-based system was simply displaying chaotic behavior. The first counter-argument had already been furnished by the clustering exercise in Fig. 5, where the system behavior was shown to be robust, without undue sensitivity to the change in the simulation time step. We would expect high sensitivity if the observed divergence was primarily due to random fluctuations. Forcing soldiers to use only warning shots was a significant change and yet the divergence persisted. We can therefore quite safely conjecture that the divergence is caused by some stable and robust mechanism. This does not mean that the element of randomness plays no role – in fact it has to because without it the simulations would be completely deterministic – but that there are likely to be other, deterministic factors amplifying and stabilizing the divergence.

Our fourth hypothesis was that the external events and social influence acted together, perhaps as part of a two-stage or even a multi-stage process. However, in order to verify this we needed to improve our method first.

## 5.1 Method Improvements and New Preliminary Results

As mentioned above, the first improvement aimed at identifying decisive moments early in the simulations. This we have solved by logging causal partitions every 2 seconds during the simulation. Later, off-line, we could then identify the moment at which the causal partitions started exhibiting increased predictive power, and which partition components were responsible.

The second improvement reflected our need for more detailed information: instead of considering the combined effect of all the external events lumped together, we recorded the effect of soldier actions separately from that of civilian actions. This meant a split of each “external event” component of our causal partitions into two. Thus, $F_E$ was split into $F_{EC}$ (civilian actions) and $F_{ES}$ (soldier actions), and $A_E$ into $A_{EC}$ (civilians) and $A_{ES}$ (soldiers). The causal partitions of anger and fear thus became $A = (A_{A0}, A_{EC}, A_{ES}, A_I, A_L)$ and $F = (F_{F0}, F_{EC}, F_{ES}, F_I)$.

At present, our exploratory analysis with the improved method proceeds along two dimensions. The first (and currently the more advanced) relates to data mining, namely to alternative ways of determining the relative importance of causal attributes for prediction, e.g. through decision trees. We have published a preliminary study [6], where we have shown that the 10th second of the simulated time was the earliest moment in which the outcome could be predicted with an increased accuracy (72%), mainly thanks to the component $F_{EC}$ (the effect of civilian actions on fear). In the 12th second, the prediction accuracy jumped to 87.6%, but here the importance of $F_{EC}$ faded, having been replaced by $F_I$ (the effect of social influence on fear), followed by $F_{ES}$ (the effect of soldier actions on fear). In the 14th and 16th seconds the prediction accuracy reached 98.6% and 99.1%, respectively, and here $F_I$ strengthened its lead, followed by $F_{EC}$ and $F_{ES}$ (in that order). Thus we indeed saw the expected staged process: the external events (civilian actions) starting it, and social influence taking over. Actions of soldiers seemed to play a temporary and intermediary role. We consider the fact that so early in the scenario we could predict so precisely its subsequent aggressive or timid turn (covering 90 seconds of simulated time) as very significant and promising. Yet, this result is not exactly what we had hoped for, because we still do not know which feature of our model is causing it. In order to answer that question, we need to devise new hypotheses and new experiments, which we envisage as the second, and more challenging, dimension of our future work.

## Conclusions

In this paper we have shown how the principles of structural causal analysis can be adapted for exploratory analysis of a hybrid dynamical system whose continuous dynamics is described by ordinary nonlinear differential equations. The key step in the process is the introduction of “causal partitions” of model variables – vector-like structures whose components quantify the influence of various causal factors on a given variable. Causal partitions can be processed by machine learning techniques and assist in the process of meaningful interpretation of the emergent behavior of the simulated system.

In our practical experiments we have demonstrated the *relevance* of causal partitions, that is, their ability to classify the simulations accurately into two classes – timid and aggressive. Regarding the *practical utility* of our method, we formulated several hypotheses and tried to qualitatively assess their likelihood based on the results of our clustering and classification experiments. We have also implemented the improvements proposed in [3] and presented the first tentative results reached with the improved method. The successful resolution of our task requires further work along two dimensions:

- To develop reliable methods of determining the relative importance of causal attributes for prediction
- To formulate new hypotheses regarding the underlying causes of the observed behavior, and design new experiments to verify them

In spite of the work that is yet to be done, we feel justified in concluding that structural causal analysis and causal partitions represent potentially valuable tools not only for hybrid dynamical systems, but also for simulation studies in general.

## Acknowledgement

This work was supported by the European Defence Agency project A-0938-RT-GC EUSAS, by the Slovak Research and Development Agency under the contract No. APVV-0233-10, and by projects KC-INTELINSYS ITMS 26240220072 and VEGA No. 2/0054/12.

## References

[1] J. Y. Halpern and J. Pearl, “Causes and Explanations: a Structural-Model Approach. Part I: Causes,” *Brit. J. Phil. Sci.*, 56, pp. 843-887, 2005

[2] J. Y. Halpern and J. Pearl, “Causes and Explanations: a Structural-Model Approach. Part II: Explanations,” *Brit. J. Phil. Sci.*, 56, pp. 889-911, 2005

[3] M. Kvassay, L. Hluchý, P. Krammer and B. Schneider, “Exploring Human Behaviour Models through Causal Summaries and Machine Learning,” in *Proceedings of INES 2013*. Budapest: IEEE Industrial Electronic Society, 2013, pp. 231-236

[4] M. Kvassay, L. Hluchý and B. Schneider, “Summarizing the Behaviour of Complex Dynamic Systems,” in *Proceedings of SAMI 2013*. Piscataway: IEEE, 2013, pp. 15-20

[5] M. Kvassay, L. Hluchý, B. Schneider and H. Bracker, “Towards Causal Analysis of Data from Human Behaviour Simulations,” in *Proceedings of LINDI 2012*. Piscataway: IEEE, 2012, pp. 41-46

[6] M. Kvassay, P. Krammer and L. Hluchý, “Validation of Parameter Importance in Data Mining: a Case Study,” in *Proceedings of WIKT 2013*, Eds. F. Babič and J. Paralič. Košice: Centre for Information Technologies, Technical University in Košice, 2013, pp. 115-120

[7] B. Latané, “Dynamic Social Impact,” in *Philosophy and Methodology of the Social Sciences*, Vol. 23, R. Hegselmann, U. Mueller and K. G. Troitsch, Eds. Dordrecht: Kluwer Academic Publishers, 1996

[8] J. Pearl, “The Art and Science of Cause and Effect” 1996 Faculty Research Lecture, http://singapore.cs.ucla.edu/LECTURE/lecture_sec1.htm

[9] J. Pearl, *Causality: Models, Reasoning, and Inference.* New York: Cambridge University Press, 2000

[10] J. Pearl, "An Introduction to Causal Inference," *The International Journal of Biostatistics* Vol. 6 (2010), Issue 2, Article 7. Available at: http://ftp.cs.ucla.edu/pub/stat_ser/r354-corrected-reprint.pdf

[11] J. Pearl, “Reasoning with Cause and Effect,” 1999 IJCAI Award Lecture, URL: http://singapore.cs.ucla.edu/IJCAI99/index.html

[12] J. Platt, “Fast Training of Support Vector Machines using Sequential Minimal Optimization,” in B. Schoelkopf and C. Burges and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, 1998

[13] B. Schmidt, “Modelling of Human Behaviour: The PECS Reference Model,” in *Proc. 14th European Simulation Symposium*, A. Verbraeck, W. Krug, Eds. SCS Europe BVBA, 2002

[14] C. Urban, “PECS A Reference Model for the Simulation of Multiagent Systems,” in *Tools and Techniques for Social Science Simulation*, R. Suleiman; K. G. Troitzsch and N. Gilbert, Eds. Heidelberg; New York: Physica-Verlag, 2000, pp. 83-114

[15] Weka 3: Data Mining Software in Java, 2013, URL: http://www.cs.waikato.ac.nz/ml/weka