Hindawi
Complexity
Volume 2017, Article ID 8381954, 18 pages
https://doi.org/10.1155/2017/8381954

Research Article

Causal Analysis of an Agent-Based Model of Human Behaviour

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

[1]*Ústav Informatiky SAV, Dúbravská Cesta 9, 84507 Bratislava, Slovakia*
[2]*Airbus Defence and Space GmbH, Rechliner Strasse, 85077 Manching, Germany*

Correspondence should be addressed to Marcel Kvassay; marcel.kvassay@savba.sk

Received 12 July 2016; Accepted 17 November 2016; Published 24 January 2017

Academic Editor: David Arroyo

This article investigates causal relationships leading to emergence in an agent-based model of human behaviour. A new method based on nonlinear structural causality is formulated and practically demonstrated. The method is based on the concept of a *causal partition* of a model variable which quantifies the contribution of various factors to its numerical value. Causal partitions make it possible to judge the relative importance of contributing factors over crucial early periods in which the emergent behaviour of a system begins to form. They can also serve as the predictors of emergence. The time-evolution of their predictive power and its distribution among their components hint at the deeper causes of emergence and the possibilities to control it.

1. Introduction

When we study a complex system whose model equations cannot be solved analytically, we typically turn to simulations. Social systems are investigated in this way too, especially their emergent behaviours. The preferred and natural approach is then agent-based modelling. In agent-based models, even relatively simple rules governing individual agents tend to produce unexpected patterns of global behaviour. The phenomenon of emergence poses two main challenges:

(i) Validation challenge: how to ascertain whether the phenomenon observed in the simulation model can also occur in the real system

(ii) Explanatory challenge: how to identify its causes and stages of manifestation

In this article, we focus on the explanatory challenge. As an aside we may mention that the two challenges are in fact interrelated since explanation is often a prerequisite to validation. Thus an explanation might reveal that a particular phenomenon was caused by the simplifications of the simulation model and should not occur in the real system (rather the simulation model should be refined). In other cases, the explanation might suggest experimental conditions under which the emergent phenomenon could be observed in the real system too. For this reason, although our proposed approach addresses the explanatory challenge, it can also indirectly serve as an auxiliary model validation tool, for example, for operational validation in the sense of Sargent [1] or Louie and Carley [2].

In general, emergent behaviour can be very difficult to explain, and we still lack powerful and general methods for the purpose. The European Agent Technology Roadmap [3] states that "understanding the mechanisms that can be used to model, assess and engineer self-organisation and emergence in multi-agent systems is an issue of major interest." Similarly, the UK governmental report [4] claims that the "difficulty in forming rigorous causal characterisations of the aggregate behaviour of a complex system (rather than the *absence* of regularity or predictability in this aggregate behaviour) …is the more legitimate barrier to adopting complex-systems approaches in an ICT engineering context."

In this article, we do *not* attempt to tackle the problem in its full generality; instead, we try to show how to build narrower methods tailored to the specifics of the system or model at hand on the basis of nonlinear structural theory of causation formulated by Judea Pearl and other researchers [5–7]. In doing so, we extend and conclude our earlier attempts to apply structural causality to an agent-based model of human behaviour [8].

Broadly speaking, if we already have a theoretical description of an agent-based model (ABM) and its executable

implementation in the form of an agent-based simulator (ABS), our approach consists of four steps:

(1) *Model analysis*: ABM equations are analysed from the point of view of structural causality in order to define additional "causal analytical variables" (CAVs) that need to be calculated and logged during the simulations. As we explain later on, each CAV quantifies the effect of one causal factor on a given model variable and represents one component of its causal partition

(2) *Implementation and data provision*: the algorithms for computing and logging the CAVs are implemented in the ABS, and appropriate simulation experiments are then defined and executed

(3) *Data analysis*: logs of the simulation runs deemed relevant for causal analysis are analysed by suitable machine learning techniques (clustering, classification, etc.) using CAVs as predictors in order to generate and test hypotheses about the causes and stages of the observed emergent behaviours

(4) *Hypothesis validation* by new simulation experiments: in this phase, we go beyond statistical hypothesis testing by taking advantage of the fact that simulation models are fully observable and manipulable by setting their initial conditions and parameters and, if necessary, by modifying their software implementation. Thus we can ultimately prove or disprove our causal hypotheses by directly manipulating the suspected causes and observing the effects of our interventions on the emergent behaviour of the simulated system.

In the remainder of this article, we walk through this process focusing mainly on steps (1), (3), and (4). We start by describing our agent-based model and scenario in Section 2. In Section 3, we introduce structural causality and the concept of a *causal partition*. In Section 4, we incrementally develop the method of causal partitions in the context of our agent-based model. This section corresponds to the "model analysis" step of the above process. In Section 5, we then analyse the results of our simulation experiments, generate hypotheses, and validate them by additional experiments.

2. The Model and the Scenario

Both the model and the scenario come from project EUSAS ("European Urban Simulation for Asymmetric Scenarios") [9] financed by 20 nations under the *Joint Investment Program Force Protection* of the European Defence Agency (EDA). The project dealt with asymmetric security threats in which security forces face rioting crowds, insurgents, or terrorists rather than regular military forces. One such example, which provided context for our simulation scenario, was the peacekeeping ISAF mission in Afghanistan. In the scenario, a crowd of civilians is looting a shop and an approaching soldier patrol is supposed to stop the looting and disperse the crowd. The scene is depicted in Figure 1. The black areas represent buildings and barriers unreachable to agents. The rectangle with gray interior near the top is the looted shop.

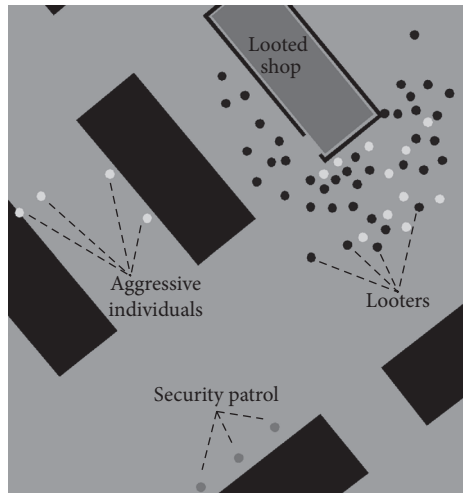

Figure 1: Initial setting of the simulation scenario.

It is surrounded by dots, each representing one agent. The dark ones are the looters; the white 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.

Civilian agents are endowed with one "default" motive and a matching behaviour by which they try to satisfy it. For looters, this leads to looting and for the violence-prone individuals to stone-pelting the soldiers. The agents also monitor their surroundings. As the patrol approaches, this may induce fear in some looters who then start leaving the scene. The violence-prone individuals, however, do not get afraid but rather 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. The ratio of looters who get afraid to those who get angry depends on their motivational dynamics, which we explain next.

A simplified diagram of the key factors affecting the behaviour of our civilian agents is shown in Figure 2. The model draws primarily on Berkowitz [10], Prentice-Dunn and Rogers [11], Staub [12], and Cañamero [13]. The main ideas and processes underlying the model were developed by Airbus Defence and Space (former EADS Deutschland GmbH) (https://airbusdefenceandspace.com/) in collaboration with the Department of Social Psychology of the University of Zurich and the chair for Operations Research at the University of Passau on behalf of the German Bundeswehr.

As depicted in Figure 2 (starting from the top left corner), the number of people surrounding the agent, their actions, and other events in the vicinity affect the agent's emotional motives *(fear, anger)* and other internal variables *(arousal, readiness for aggression)*. Besides events and actions, there is also a direct *social influence* of other agents on the agent's *fear* and *anger*. This was modelled according to Latané's formula of strength, physical proximity, and the number of influencing agents [14].

Speaking qualitatively, the agent's internal *arousal* depends on the number of people in the vicinity and their

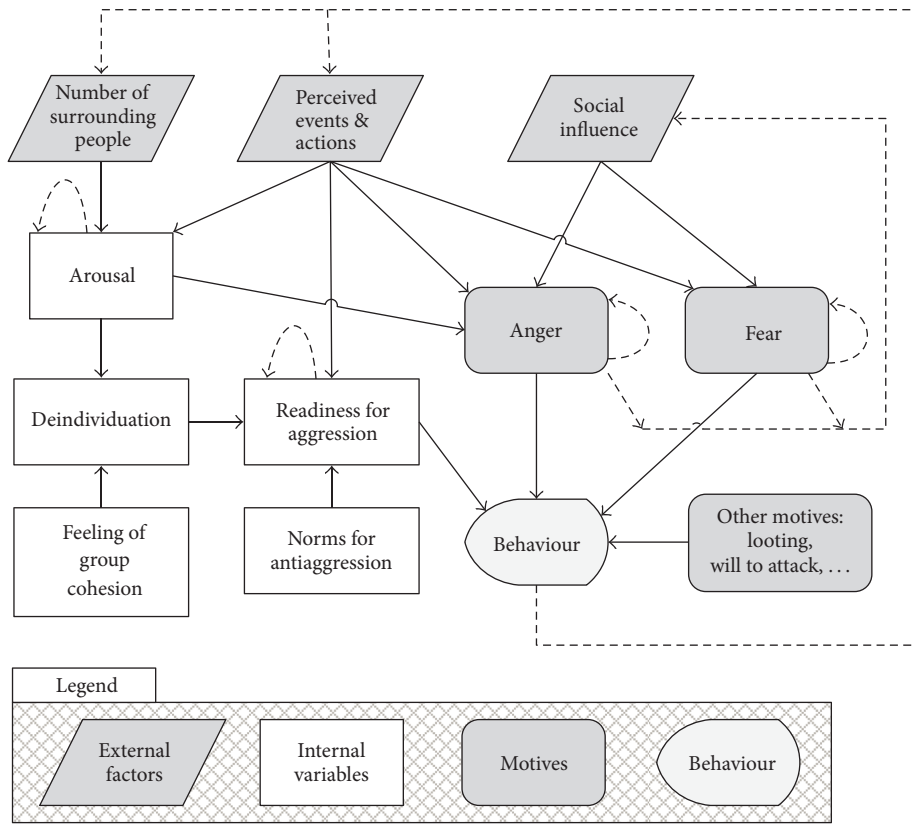

Figure 2: Key factors affecting the behaviour of civilian agents.

violence: the higher the number and the more violent they are, the sharper the increase of the agent's arousal. *Deindividuation* means that the agent considers himself a part of the crowd and no longer a separate individual: the higher the agent's arousal and the cohesion of his group, the higher the deindividuation. *Readiness for aggression (RFA)* is jointly affected by the norms for antiaggression, deindividuation, and external events as follows: (a) the higher the norms for antiaggression, the lower the *RFA*; (b) the higher the deindividuation, the higher the *RFA*; and (c) the more violent actions are witnessed, the higher the RFA is. The form of the agent's eventual aggression depends primarily on the *RFA*. In our scenario, the initial value of *RFA* was set to a level guaranteeing that civilians would resort to stone-pelting the soldiers whenever they became aggressive.

Regarding the agent-based architecture suitable for implementation, project EUSAS opted for the PECS reference model [15, 16]. The "PECS" acronym stands for *P*hysical conditions, *E*motional state, *C*ognitive capabilities, and *S*ocial status, the four kinds of internal factors that need to be modelled in order to achieve realistic agent behaviour. At the same time, the PECS model only provides these four empty slots without specifying the factors to be modelled or the level of modelling detail: these decisions are left to the modeller as they depend on the task at hand. The PECS simply requires that these factors be modelled as state variables with associated state transition functions conforming to general systems theory.

The PECS also requires that some state variables produce or act as *motives*, that is, forces that drive agents to action. A good example is the level of physical energy (a state variable in the "physical conditions" slot) and hunger, a *motive* driving us to search for food and replenish the energy. Our model includes four state variables that act directly as motives: *fear, anger, looting* motive (present only in looters), and *will to attack* (present only in violence-prone civilians). In the PECS model, the motives compete for control over the behaviour of the agent: the strongest wins and becomes *action-guiding*. The motives thus need to be mutually comparable, which we have achieved by normalizing and restricting their values to the closed interval $[0, 1]$. Unlike *fear* and *anger*, which were endowed with complex dynamics described below, the *looting* motive and *will to attack* were both set to constant values which *fear* and *anger* had to cross in order to affect the agents' behaviour.

Each motive, when it becomes action-guiding, preselects a group of behaviours that can satisfy it. In our model, for example, there are three "fearful" behaviours: *withdrawal* (walking away), *flight* (running away), and *panic flight* (running away at extra speed with sensory perception blocked). Which of them is triggered when *fear* becomes action-guiding depends on a secondary selection criterion, in this case the actual intensity of *fear*. This criterion can

Table 1: Fear- and anger-related constants.

Constant	Value	Interpretation
c_{1F}	5	Sensitivity of the first derivative of fear dF/dt.
c_{2F}	1.1	Maximum value of fear F at which dF/dt becomes zero.
c_{3F}	0.1	Inbuilt tendency of fear to increase, a composite constant defined as the difference between the effect of the expected negative consequences and the agent's resilience to fear.
c_{4F}	12.5	Sensitivity to fear-related social influence of other agents.
c_{5F}	1	Sensitivity to fear-inducing events (see Table 2).
c_{1A}	2	Sensitivity of the first derivative of anger dA/dt.
c_{2A}	1.1	Maximum value of anger A at which dA/dt becomes zero.
c_{3A}	0.3	Resilience to anger.
c_{4A}	12.5	Sensitivity to anger-related social influence of other agents.
c_{5A}	1	Sensitivity to anger-inducing events (see Table 2).

be arbitrary; for example, while *anger* preselects a group of aggressive behaviours, the final choice of the form of aggression depends on the *readiness for aggression (RFA)*. As we have already mentioned, we set *RFA* so that aggressive civilians would always resort to stone-pelting the soldiers.

In line with the PECS modelling methodology, agent behaviours are conceptualised as sequences of atomic, uninterruptible *elementary actions*, for example, one step in a certain direction or one stone-throw. When a new motive becomes action-guiding, it only takes effect after the current elementary action is completed: the current behaviour pattern is then cancelled and a new one is activated.

As for the dynamics of the simulated emotions *fear* and *anger*, they comprise a continuous part and a discrete part. The continuous dynamics of *fear* (F) is driven by the differential equation

$$\frac{dF}{dt} = c_{1F} \cdot F \cdot (c_{2F} - F) \cdot (c_{3F} + I_F), \tag{1}$$

where c_{1F}, c_{2F}, c_{3F} are fear-related constants (see Table 1) and I_F is fear-related social influence of nearby agents. Analogously, the continuous dynamics of *anger* (A) is driven by the equation

$$\frac{dA}{dt} = c_{1A} \cdot A \cdot (c_{2A} - A) \cdot (I_A + L - c_{3A}), \tag{2}$$

where c_{1A}, c_{2A}, c_{3A} are anger-related constants (see Table 1), I_A is anger-related social influence of nearby agents, and L is the model variable *arousal*.

Social influence I_V of nearby agents on a motive variable V (where V stands for either *fear* F or *anger* A) of an observing agent j is defined by the following sum:

$$I_V = c_{4V} \cdot \sum_{k \neq j} \left[\frac{(V_k - V_j) \cdot \text{prestige}_k \cdot \text{sympathy}_{jk}}{\text{distance}_{jk}} \right]. \tag{3}$$

Here, the summation is over those agents (indexed by the subscript k) who are not farther away from agent j than a certain *social influence radius* (set to 100 metres for both *fear* and *anger*) and in whom the motive V happens to be *action-guiding* at the moment of evaluation. Each agent is assigned a constant social rank prestige_k which modulates its influence on others: group leaders enjoy higher prestige than ordinary members and thus influence the others more. Analogously, there are constant sympathies assigned between groups: sympathy_{jk} captures the sympathy of agent j's group toward agent k's group and is interpreted here as the susceptibility of the former to the influence of the latter. Finally, distance_{jk} is the physical distance between the two agents. In the original model, variables and many constants were expressed in the percentage scale $[0, 100]$. For the sake of simplicity, in this paper, we have converted them into the ratio scale $[0, 1]$ (see Table 1).

During the simulation, differential equations (1) and (2) for each agent are solved numerically by the Euler method with a constant (but user-definable) time step Δt. Resorting to numerical method enabled us to ignore the specific form of the equations' right-hand side and consider the general case

$$\frac{dF}{dt} = f(F, I_F), \tag{4a}$$

$$\frac{dA}{dt} = h(A, I_A, L), \tag{4b}$$

where f, h are bounded (but not necessarily continuous) nonlinear functions. The Euler method approximates the new values of fear $F(t + \Delta t)$ and anger $A(t + \Delta t)$ on the basis of the current ones $F(t)$, $A(t)$:

$$F(t + \Delta t) \approx F(t) + \Delta t \cdot f(F(t), I_F(t)), \tag{5a}$$

$$A(t + \Delta t) \approx A(t) + \Delta t \cdot h(A(t), I_A(t), L(t)). \tag{5b}$$

After calculating these new "continuous" values, discrete dynamics come into play: the cumulative effects of the perceived external events on *fear* (ΔE_F) and *anger* (ΔE_A) are added in order to obtain the new "total" values:

$$F_T(t + \Delta t) = F(t + \Delta t) + c_{5F} \cdot \Delta E_F, \tag{6a}$$

$$A_T(t + \Delta t) = A(t + \Delta t) + c_{5A} \cdot \Delta E_A. \tag{6b}$$

In the next iteration, the new total values of *fear* F_T and *anger* A_T will be used as initial conditions in the numerical solution of the differential equations representing their continuous dynamics. This sequential coupling of the continuous and the discrete dynamics qualifies our agent models as *sequential hybrid* in the sense of Swinerd and McNaught [17].

Table 2: Main event impacts on fear and anger.

Event	Impact on fear		Impact on anger	
	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 constants c_{5F}, c_{5A} (see Table 1) capture the agent's individual sensitivity, while ΔE_F and ΔE_A are the sums of the emotion-inducing impacts as per Table 2 for all the events perceived by the agent during the time interval $(t, t + \Delta t)$. The "direct" values from Table 2 are used when the perceiving agent is within 20% of the maximum perception distance. When the agent is farther away than 40% of this maximum distance, the "indirect" values are used. In the intermediate zone (from 20% to 40%), a weighted average is used, sliding down linearly from the direct value toward the indirect one. Sensory perception of the agents is limited by a radius of 50 m for events like throwing stones and 150 m for gun shots.

Besides the constants in Table 1 and the event impacts in Table 2, emotional dynamics is greatly influenced by the initial values of *fear* F_0 and *anger* A_0. In our scenario, these were set to $F_0 = 0.3$ and $A_0 = 0.2$. Additionally, our agents underwent the moderating influence of *emotions* and *fatigue*: when the average of *fear* F and *anger* A (termed in our model the *emotional moderator*) crossed the level of 0.5, further sensory perception of external events was blocked. Moreover, as the physical energy of our agents decreased with expended effort and sustained injuries, they slowed down and their actions took longer.

Like all models, this one too is no more than a "useful simplification" of the enormously complex human psyche. Its intended use was for virtual training of security personnel in project EUSAS; therefore its equations were formulated in a deterministic fashion so as to lead to reasonably predictable agent behaviour. Some measure of variation in its behaviour was subsequently achieved by randomizing agents' initial positions and the duration of their elementary actions.

Practitioners of System Dynamics (SD) will have undoubtedly noticed that we have modelled agent motives by the same kind of equations as those used in SD. Here, of course, the context is different: our equations describe an "inside" of a single agent, whereas in SD such equations would typically be used to model the whole agent community without bothering to model its constituent individuals or their low-level interactions. Despite this difference in modelling focus, the similarity of the equations used and the emphasis that SD puts on tracing the flow of causality through the modelled system make it very likely that our method of causal partitions could be profitably employed in SD context. In Conclusion, we therefore provide some practical hints in this regard.

Numerical solution of this model by iterating through (5a) and (5b) and (6a) and (6b) can give us a complete and detailed time-evolution of simulated fear and anger of civilian agents. Our present goal, however, is more ambitious. In the case of fear, for example, we want to know what portion of its actual value at any time should be attributed to the social influence of nearby agents (variable I_F in (5a)) as opposed to the direct impact of external events (variable ΔE_F in (6a)). We tackle this question in the sections that follow.

In contrast to our civilian agents, the soldier patrol characters were intended primarily as avatars to be controlled by real people in a high-fidelity 3D cyber-environment of a commercial battlefield simulator VBS2 (http://www.army-technology.com/contractors/training/bohemia-interactive/). In consequence, our soldier agents were not defined at the same level of detail as the civilian ones. In our scenario, for example, they are just passing by and act in self-defence. Their rules of self-defence say that when a given civilian first throws a stone at a particular 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, that 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.

While experimenting with this model and scenario, we noticed that, for the parameter setting listed above, the emergent collective behaviour of the civilians seemed to bifurcate along two different trajectories. In some cases, almost all the looters got afraid and left the scene, while in others almost all got angry and joined the attack. Because our model incorporated an element of randomness, some variation in its behaviour was expected, but the extreme variation we witnessed was unusual and called for an explanation. This spurred our search for analytical methods that could unravel causal chains and dependency in complex systems of this kind. The method of *causal partitions* presented below is the result.

3. Structural Causality and Causal Partitioning

Structural causality starts from the concept of a structural equation. In order to illustrate it, let us consider two simple electrical circuits shown in Figure 3. Both include a variable resistor connected to an ideal source of (a) voltage or (b) current.

The relationship between the current I passing through the resistor and the voltage U on its terminals depends on its resistance R and conforms to Ohm's law, which can be expressed in many nearly equivalent ways:

$$\begin{aligned} I \cdot R &= U; \\ \frac{U}{I} &= R; \\ I &= \frac{U}{R}; \\ R &= \frac{U}{I}. \end{aligned} \tag{7}$$

From the algebraic point of view, all these formulations are permissible and any of them could be said to apply to both circuits. Structural causal theory, by adding extra

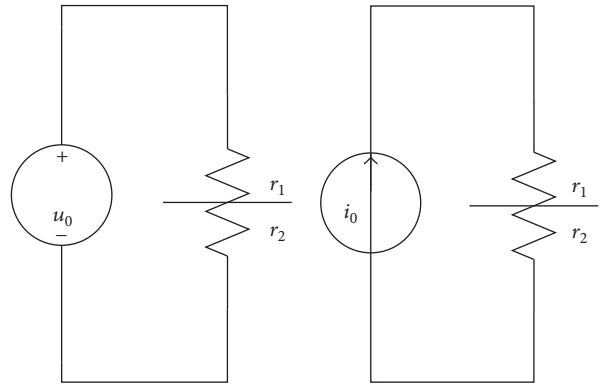

Figure 3: Two simple electrical circuits.

rules governing the form of equations, manages to encode in them additional information about the flow of causality. Essentially, it treats the equality sign as an assignment operator in programming languages. Thus, on the left-hand side of a structural equation, there can only be one variable. Moreover, this variable has to be genuinely "dependent" on the right-hand side, which is meant to capture the causal mechanism determining (or "assigning") its value in the system under investigation. Interpreted in this way, each of the two circuits can only be represented by one form of Ohm's law. In order to identify the correct "structural" forms, we need to contemplate the effect of an "intervention" by an experimenter: what happens if he or she changes the value of the resistor from r_1 to r_2? We can immediately see that for (a) it is the current that changes, while for (b) it is voltage. Thus the right structural equations describing the circuits are (a) $I = U/R$ and (b) $U = I \cdot R$.

Let us now consider a slightly modified version of Figure 3(b): instead of an ideal current source we might have a photovoltaic cell producing current dependent on the intensity of light (E), and our resistor might be a thermistor whose resistance depends on temperature (T). A simplified structural model of such a circuit might look as follows:

$$I = g(E), \tag{8a}$$

$$R = h(T), \tag{8b}$$

$$U = I \cdot R, \tag{8c}$$

where g, h are unspecified nonlinear functions. By including only these equations in the model we stipulate that there is no other significant dependency among the model variables, at least in the space of parameter values under consideration. Thus, for example, we assume that the current I produced by the photovoltaic cell does not depend on the resistance R, a condition which probably holds just approximately and only for a certain range of resistance values. We also assume that the temperature T of our thermistor does not depend either on the intensity of light E or on the current I passing through it; otherwise we should have represented this dependence explicitly by a dedicated equation. It is not our intention to defend this model as realistic; it serves merely to illustrate the principles of structural causality and our method of causal partitions.

The variables that occur on the left-hand side of structural equations are called *endogenous*, that is, determined by the model. In our example, these are $\{I, R, U\}$. The remaining variables $\{E, T\}$ are *exogenous*: in their case, we are not interested in modelling the mechanisms that set their values and simply consider their values as given. The value assignment to the exogenous variables, for example $(E = e_1, T = \tau_1)$, is called the *background* or *context* in which we try to solve the model equations.

In general, according to Pearl [5], causal analysis can be applied to systems that are described by structural equations of the form

$$X_k = f_k(pa_k, u_k), \quad k = 1, \ldots, n, \tag{9}$$

where pa_k stands for the set of "parent variables" of X_k (i.e., endogenous variables directly determining the value of X_k through an autonomous causal mechanism captured by f_k) and u_k represents the set of background variables. The autonomy of the mechanisms (and, consequently, of the equations) means that it should be possible to change any of them by external intervention without affecting the remaining ones [18]. A set of such equations is called a "structural model."

In structural theory, causation is interpreted as a relation between *events*. A *primitive event* is defined as a model variable assuming a value from its permitted range; for example, $R = r_1$ [6]. More complex events can be expressed as Boolean combinations of primitive events. We normally speak of an *event* when something happens as a consequence of model equations; for example, in our case, the *event* $R = r_1$ would imply that the temperature T assumed a value τ_1 such that $h(\tau_1)$ is evaluated to r_1. Causal thinking also requires a special kind of event called *intervention* (sometimes also *action*): this is when we intervene from "outside" and impose the value assignment $R = r_1$ regardless of the temperature or the causal mechanism $h(T)$, for example, by replacing the thermistor with a normal resistor whose resistance equals r_1. An intervention means that we wipe out the affected

structural equation from the model and replace it with another (typically a straightforward value substitution). The solution of this modified system of equations represents the response of the model to the intervention. In our case, (8b) would be replaced with $R = r_1$, giving the model response $U = g(e_1) \cdot r_1$.

Structural approach to causality enables us to inquire whether one event ($X = x$) is a cause of another ($Y = y$) in a given context $C = c$. Various kinds of causes have been proposed in the literature, but we shall focus here on the notion of *actual cause* as defined by Halpern and Pearl in [6]. Informally, $X = x$ is an actual cause of $Y = y$ if the following conditions hold (in general, X, Y, C can represent subsets of model variables. If $X = \{X_k \mid k = 1, \ldots, n\}$, then $X = x$ denotes the logical conjunction of primitive events $\bigwedge_{k=1}^{n} X_k = x_k$):

(i) *Actuality*: both $X = x$ and $Y = y$ are true (observed) in the model in the context $C = c$

(ii) *Dependence*: an intervention on X changing its value from x to some other x' ($x' \neq x$) must result in a change of the value of Y from y to some other y' ($y' \neq y$). This needs to be demonstrated for at least one setting $W = w'$ of a suitable subset of the remaining endogenous variables, which is to be imposed through *intervention*

(iii) *Sustenance*: the intervention $X = x$ must insulate the event $Y = y$ from a restricted class of interventions on the value of W. Specifically, if any subset of the variables in W is made to flip between the values w' and w (w being the original value of W observed under the context $C = c$ before the intervention $W = w'$), this must not have any effect on the value of Y, so long as X remains set to x

(iv) *Minimality*: X is minimal; no proper subset of X satisfies the above conditions

This informal rendering of Halpern and Pearl's definition is not entirely precise but suffices for our purposes (for rigorous definition, please refer to [6] or [5]).

Armed with this definition, we can analyse our circuit model. For example, we can inquire whether the event $R = r_1 = h(\tau_1)$ qualifies as an actual cause of event $U = g(e_1) \cdot h(\tau_1)$ observed at time t_1 in the context $C = c_1 = (E = e_1, T = \tau_1)$. The first condition, *actuality*, is satisfied, since both events hold. In other words, we obtain them as a solution to the model equations ((8a), (8b), and (8c)) in this context. Assuming that all our variables are real-valued and positive, *dependence* is also easily demonstrated: it is trivial to show that, for example, halving the value of R would halve the value of U. The fact that we could demonstrate this without intervening on the remaining endogenous variable I makes the *sustenance* condition also trivially satisfied: if we include I in W, then the value w (corresponding to $I = i_1 = g(e_1)$ observed in the context $C = c_1$) is the same as the value w' for which *dependence* was demonstrated. Flipping between w and w' then means that nothing actually changes, so the observed value of $U = g(e_1) \cdot h(\tau_1)$ cannot change either. Thus, *sustenance* is (trivially) satisfied too. Last, $R = r_1 = h(\tau_1)$ is *minimal* since it involves only one model variable. Therefore, it qualifies as the actual cause. Moreover, we could have equally easily demonstrated that $I = i_1 = g(e_1)$ also is an actual cause of $U = g(e_1) \cdot h(\tau_1)$. This ease amounting to triviality in fact signals a problem: applying structural causality in this form to models with real-valued variables is not going to be very useful or instructive. Typically, all the parent variables will be identified as actual causes of their dependent or "child" variable assuming its observed value: a result no doubt correct but not particularly enlightening. This stands in sharp contrast to highly nontrivial results obtained in models with binary or categorical variables, as documented by the many interesting examples in [6].

In our previous work [8], we suggested how to extend the structural approach in order to cope with real-valued variables. Instead of asking, "What is the cause?" we proposed to ask a modified question: "In what proportion have all the causes contributed to the effect?" Continuing with our example, we should try to determine the proportion in which the change of U from some previous level u_0 to the present one u_1 could be attributed to (or split among) its parent variables I and R. Let us assume that $U = u_0$ was observed at time t_0 together with $I = i_0, R = r_0$ in the context $C = c_0 = (E = e_0, T = \tau_0)$. For the sake of simplicity, let us further assume that our two observations were so close in time that the state-space trajectory of the system between them can be considered linear. In such a case, we can approximate the change of U by a scalar product of two vectors. The first is the gradient of U along its parent variables according to its structural equation (8c). The second is the vector form of the change in these parent variables between the two observations (contexts $C = c_0$ and $C = c_1$):

$$\Delta U \approx \operatorname{grad}(U) \cdot (\Delta I, \Delta R) \tag{10a}$$

$$\approx \left(\frac{\partial U}{\partial I}, \frac{\partial U}{\partial R} \right) \cdot (\Delta I, \Delta R) \tag{10b}$$

$$\approx \frac{\partial U}{\partial I} \cdot \Delta I + \frac{\partial U}{\partial R} \cdot \Delta R. \tag{10c}$$

Equation (10c) can be interpreted as a general "recipe" for quantifying the responsibility of the parent variables I, R for the change of their dependent variable U: the first summand on the right-hand side ($\Delta I \cdot \partial U / \partial I$) captures the contribution of I and the second one ($\Delta R \cdot \partial U / \partial R$) captures that of R.

Suppose that we trace further evolution of U at times $t_2, \ldots, t_m$ as it assumes new values $u_2, \ldots, u_m$ along some trajectory representing our observational study or experiment. Expressing each of these changes of U as per (10c) and summing up the contributions of I separately from those of R, we obtain a general formula quantifying the total contribution of each causal factor (parent variable) along an arbitrary trajectory (implicitly approximated here by a piecewise linear curve):

$$\sum_{j=1}^{m}\Delta U^{(j)} \approx \sum_{j=1}^{m}\left(\Delta I^{(j)}\cdot\left.\frac{\partial U}{\partial I}\right|_{\substack{I=\widehat{I}^{(j)}\\R=\widehat{R}^{(j)}}}\right) + \sum_{j=1}^{m}\left(\Delta R^{(j)}\cdot\left.\frac{\partial U}{\partial R}\right|_{\substack{I=\widehat{I}^{(j)}\\R=\widehat{R}^{(j)}}}\right). \tag{11}$$

The above formula uses superscript indexing, where $X^{(j)}$ denotes the value of variable X at time t_j and $\Delta X^{(j)}$ denotes its backward difference: $\Delta X^{(j)} = X^{(j)} - X^{(j-1)}$. Partial derivatives are evaluated at midpoint $(\widehat{I}^{(j)}, \widehat{R}^{(j)})$ of each linear segment of the trajectory, so $\widehat{I}^{(j)} = 0.5\cdot(I^{(j)} + I^{(j-1)})$ and $\widehat{R}^{(j)} = 0.5\cdot(R^{(j)} + R^{(j-1)})$.

In order to keep the total contribution of each causal factor separate, we proposed a new, vector-like representation of model variables termed a *causal partition*. Let us consider variable Y driven by structural equation $Y = f(X_k,\ k = 1, \ldots, n)$, where $\{X_k,\ k = 1, \ldots, n\}$ are its parent variables. Let us assume that Y starts from $Y = y_0$ and proceeds through $Y = y_j,\ j = 1, \ldots, m$. Its *causal partition* vector at the end of this trajectory ($Y = y_m$) is then expressed as $(Y_{Y0}, Y_{X1}, Y_{X2}, \ldots, Y_{Xn})$, where each partition component Y_{Xk} stands for the total contribution of the corresponding parent variable X_k along this trajectory and is calculated as

$$Y_{Xk} = \sum_{j=1}^{m}\left(\Delta X_k^{(j)}\cdot\left.\frac{\partial Y}{\partial X_k}\right|_{\substack{X_1=\widehat{X}_1^{(j)}\\ \vdots \\ X_n=\widehat{X}_n^{(j)}}}\right) = \sum_{j=1}^{m}\left(\Delta X_k^{(j)}\cdot\left.\frac{\partial f}{\partial X_k}\right|_{\substack{X_1=\widehat{X}_1^{(j)}\\ \vdots \\ X_n=\widehat{X}_n^{(j)}}}\right), \tag{12}$$

where $(\widehat{X}_1^{(j)}, \ldots, \widehat{X}_n^{(j)})$ is the midpoint of the jth linear segment of the trajectory.

The first partition component Y_{Y0} has a special role: it denotes the contribution of the initial setting $Y = y_0$ to the final value $Y = y_m$. We introduced it in order to force the partition components to sum up to the value of the represented variable, which we found very helpful for interpreting causal partitions:

$$y_m = Y_{Y0} + \sum_{k=1}^{n}\left(Y_{Xk}\right). \tag{13}$$

This approach implies that, at the beginning of the trajectory, Y is represented by the causal partition vector $(y_0, 0, 0, \ldots, 0)$.

Going back to the example, variable U at the end of our observational study would be represented by the causal partition vector:

$$\left(u_0, \sum_{j=1}^{m}\left(\Delta I^{(j)}\cdot\left.\frac{\partial U}{\partial I}\right|_{\substack{I=\widehat{I}^{(j)}\\R=\widehat{R}^{(j)}}}\right), \sum_{j=1}^{m}\left(\Delta R^{(j)}\cdot\left.\frac{\partial U}{\partial R}\right|_{\substack{I=\widehat{I}^{(j)}\\R=\widehat{R}^{(j)}}}\right)\right). \tag{14}$$

Note that in this example the contribution of the initial value u_0 remained constant throughout the trajectory; that is, $U_{U0}^{(j)} = u_0,\ j = 1, \ldots, m$. In general, however, this contribution can vary, as we shall see for simulated fear F in our agent-based model of human behaviour.

4. Causal Partitioning in Practice

The notion of a *causal partition* of a model variable introduced in the previous section is the core of our new analytical method. Its development was in fact a long-term effort reported step by step in several publications, of which the most recent one is [8]. We present the derivation comprehensively here along with some as yet unpublished improvements and results.

4.1. Basic Version of Causal Partitioning. Before applying structural causality to our human behaviour model we need to ascertain that it meets the criteria set for structural models. In Appendix A, we demonstrate that when ordinary nonlinear differential equations used in our model are converted into difference equations (so as to solve them numerically through simulation), they at the same time become structural.

Let us now focus on the general form of differential equation driving the continuous dynamics of fear F of our agents (4a). As explained with regard to (10a), (10b), and (10c) in Section 3, function f on the right-hand side of (4a) can be linearized on each discretized time interval through the scalar product of its gradient $(\partial f/\partial F, \partial f/\partial I_F)$ and the vector form of the change in its parameters $(\Delta F, \Delta I_F)$:

$$\Delta f \approx \frac{\partial f}{\partial F}\cdot\Delta F + \frac{\partial f}{\partial I_F}\cdot\Delta I_F, \tag{15}$$

where $\Delta f, \Delta F, \Delta I_F$ stand, respectively, for the difference in the values of f, F, I_F between the start and the end of the time interval under consideration. This equation enables us to determine the proportion in which the change in the value of f can be attributed to the influence of the changes in its parameters F, I_F, which we interpret as their "elementary causal effect" on f. By summing up these elementary causal effects separately for each factor, we can partition the value of f at any moment into a sum of "total contributions" per factor:

$$f = f_F + f_S, \tag{16}$$

where f_F is the total contribution of *fear* F to the value of f (summed since the start of the simulation) and f_S is the total contribution of social influence I_F. Substituting this partition into (5a), we get

$$F(t + \Delta t) \approx F(t) + f_F\cdot\Delta t + f_S\cdot\Delta t. \tag{17}$$

If we further substitute (17) into (6a), we get

$$F_T(t+\Delta t) \approx F(t) + f_F \cdot \Delta t + f_S \cdot \Delta t + c_{5F} \cdot \Delta E_F \quad (18)$$

which can be rewritten as

$$F_T(t+\Delta t) \approx F(t) + \Delta F_F + \Delta F_S + \Delta F_E. \quad (19)$$

This means the new total value of fear can be interpreted as the old one plus the contributions of its three "causal" factors. By summing up these contributions separately since the start of the simulation, we can, analogously to (16), partition the value of F at any moment into a sum of "total contributions" per factor:

$$F = F_F + F_S + F_E. \quad (20)$$

The right-hand side of (20), written in a "vector-like" form (F_F, F_S, F_E), represents the basic causal partition of fear as per the original version of our method introduced in [19]. We interpreted its first component F_F as the extent to which fear could be considered "self-propelled."

4.2. Improved Versions of Causal Partitioning. Having worked with the basic version of our method for some time, we realised that the concept of self-propelling behaviour was problematic and complicated the interpretation of causal partitions. Eventually, we found a way to eliminate it. We present the derivation of this "enhanced" version in Section B.1 of Appendix B. In this version, the first derivative of fear f is represented by the causal partition vector (f_E, f_S) and fear F by the partition (F_E, F_S). Partition components indexed by "E" represent the contribution of external events, while those indexed by "S" stand for the contribution of social influence.

4.2.1. Dependency among Causal Factors. So far we have treated fear F and social influence I_F as if they were independent. This assumption was in fact implied in the way we used (15) for separating the effect of fear F on f from that of I_F. In consequence, both the basic and the enhanced methods silently stipulate that a change of fear F has no effect on social influence I_F. But our model does not meet this requirement: by formula (3), I_F is in fact a function of F and of fear of all the influencing agents (note that fear F corresponds to variable V_j in (3)). Because all the terms on the right-hand side of (3) are nonnegative, increasing F will simultaneously decrease I_F. In order to account for this indirect effect of F on f through I_F, the notion of a total derivative needs to be employed:

$$\frac{df}{dF} = \frac{\partial f}{\partial F} + \frac{\partial f}{\partial I_F}\frac{dI_F}{dF}. \quad (21)$$

This total derivative represents the true (or total) sensitivity of f to changes in F regardless of whether they are direct (the first term on the right-hand side) or indirect (the second term). Based on this insight we can further improve our method; the corresponding mathematical derivation is presented in Section B.2 of Appendix B.

4.2.2. Nonzero Initial Conditions and Different Types of Events. In all the versions of our method introduced so far we have implicitly assumed zero initial value of fear ($F_0 = F_T^{(0)} = 0$). In our simulation experiments, which we present in the next section, this initial value was positive ($F_0 > 0$) and strongly influenced simulation results. Its causal effects therefore need to be taken into account. As explained in Section 3, this can be achieved through a dedicated partition component, which we propose to denote by F_{F0} for fear and f_{F0} for its first derivative. The hybrid nature of our model (i.e., its combination of the continuous and the discrete dynamics) enables us to view the setting of the initial value of fear as a special kind of event which takes effect at time $t_0 = 0$ just before the simulation starts. This amounts to conceptualising the components f_{F0}, F_{F0} as being "split off" from the regular event components f_E, F_E.

As with the initialising event, we can also differentiate other kinds of events. For example, in our simulation experiments, we distinguish between the events (actions) of civilians and the events (actions) of security forces. This is easily possible since the cumulative event impact on fear ΔE_F is defined as a straightforward sum of individual event impacts which do not interact in any way:

$$\Delta E_F = \sum_k \text{impact}_F(e_k), \quad (22)$$

where impact_F is a function mapping each event e_k perceived by the agent during a given simulation step to its numerical impact on fear F. This function is defined in a tabular form in Table 2. Events are results of agent actions, and for each action we know whether its originator is a civilian or a security staff, so we can sum the impact of civilian actions (ΔE_{FC}) separately from that of security actions (ΔE_{FS}). If we then allocate a dedicated partition component to each type of event, we end up with fear F being represented by a causal partition $(F_{F0}, F_{EC}, F_{ES}, F_S)$ and its first time derivative f by a causal partition $(f_{F0}, f_{EC}, f_{ES}, f_S)$. The full derivation of recurrence relations for each partition component is presented in Section B.3 of Appendix B.

At this point it should be clear that the number of causal partition components is not fixed in advance but rather depends on the problem at hand. As a matter of fact the above structure turned out to be sufficient for our purposes, but this was by no means guaranteed. Rather, it was a matter of trial and error. If it were to be found insufficient, we would have tried to get additional information by further splits. For example, the civilians in our scenario were of several types, so we might have tried to separate the event impact for each civilian type. This would amount to splitting f_{EC} into $f_{EC1}, f_{EC2}, \ldots$ and F_{EC} into $F_{EC1}, F_{EC2}, \ldots$ as necessary. We might have also likewise split social influence I_F: individual contributions in its formula (3) do not interact in any way (they are just summed up), so the splitting would again be fairly straightforward.

This concludes our theoretical development of causal partitioning in the context of the simulated emotion of fear F of our civilian agents. It enables us to express F at any moment as a vector-like structure $(F_{F0}, F_{EC}, F_{ES}, F_S)$ whose

Table 3: Predictive accuracy of partition components and MoE (in %) at the 90th second of simulated time.

Data set	A_{A0}	A_E	A_S	A_L	F_{F0}	F_E	F_S	N_E	N_W	N_S
100 ms	88.7	61.0	93.0	51.7	99.0	97.7	80.3	97.0	95.3	88.7
300 ms	80.0	67.7	48.7	60.3	97.0	93.0	91.0	98.0	93.0	92.0

components sum up to F while at the same time isolating the portions of its value attributable to each causal factor.

Along similar lines we have developed and implemented a causal partitioning procedure for the simulated emotion of anger of our agents. Their anger A is driven by (2) and the procedure partitions it into a vector-like structure $(A_{A0}, A_{EC}, A_{ES}, A_S, A_L)$. Since the equation driving anger is similar to that driving fear, the only really new partition component is A_L, which captures the effect of a new variable L (termed *Arousal*) in (2). At this point, then, we have at our disposal two causal partitions through which we can analyse and interpret the emergent behaviour of our agent-based model. This analysis, however, requires real simulation experiments and data, which we present below.

5. Experimentation and Validation of Hypotheses

This section describes the application of our method to the data from simulations of our agent-based model and scenario. We start by briefly summarizing our early experiments and hypotheses in Section 5.1. These were already reported in detail in [8]. Practical experience gained in this early phase enabled us to improve our method, rerun the simulations, and ultimately identify the cause of the surprising emergent behaviour of our model. We present these later activities in Section 5.2. All data mining and machine learning experiments were performed in Weka [20], version 3.7.9, which used EM (expectation maximization) algorithm.

5.1. Early Experiments and Hypotheses. The set of early experiments consisted of 300 runs of our scenario with the time step $\Delta t = 300$ ms and 300 runs with the time step $\Delta t = 100$ ms. This enabled us to gauge the effect of the time step size (and of the resulting discretization and rounding errors) on the observed emergent behaviour. Since the time-evolution of our scenario was 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 were passed on to machine learning algorithms for further analysis.

In this early phase, we have not yet distinguished between the civilian and the security event impacts, so the final average values of fear F and anger A were represented by the partitions

$$F = (F_{F0}, F_E, F_S), \quad A = (A_{A0}, A_E, A_S, A_L). \tag{23}$$

Partition components F_E, A_E recorded the effect of all external events apart from the initialising one, whose effect was tracked by F_{F0}, A_{A0}. We also calculated five other relevant attributes:

(i) N_E (number of effective shots)

(ii) N_W (number of warning shots)

(iii) N_S (number of stones thrown)

(iv) *A-count* (number of times that anger became action-guiding in some civilian agent)

(v) *F-count* (number of times that fear became action-guiding in some civilian agent)

N_E, N_W, N_S are so-called measures of effectiveness (MoE) that were used to evaluate scenarios in project EUSAS: for scenarios that turned aggressive we expected high MoE and high *A-count*, while for the "timid" ones we expected low MoE and high *F-count*.

We included MoE as a sort of "competition" to our causal partitions. It was evident that MoE could classify the scenarios well, since aggressive developments implied high numbers of stones thrown as well as of gunshots. MoE, however, lacked the explanatory power: they could not tell us anything about why a particular scenario turned aggressive or timid.

An initial clustering exercise in two-dimensional space *(A-count, F-count)* revealed two distinct clusters ("timid" versus "aggressive") as expected. The shape of the clusters did not appreciably depend on the length of the time step Δt, which allowed us to rule out the discretization and rounding errors as a significant factor. The clustering supplemented our data with a new attribute assigning each simulation into one of the two clusters.

We then trained several SVM classification models on the scaled data from our data sets, choosing the cluster assignment as the target class. When we examined the predictive power of individual partition components, we were surprised to find that F_{F0} possessed the highest prediction accuracy (see Table 3). It was tempting to postulate the initial value of fear F_0 as the underlying cause, but we knew this could not be, since we had kept it constant for all the 600 simulations. Its predictive power must have stemmed from other factors, most likely the other two partition components to which it was tied by the constraint $F = F_{F0} + F_E + F_S$. Judging by their accuracy, F_E appeared to be more significant than F_S.

High predictive accuracy of these simple models proved that causal partitions were *relevant* to the investigated emergent phenomenon. The question of their *practical utility* remained unresolved, because by *practical utility* we meant their ability to guide us toward that aspect of the model which, if modified, would suppress the bifurcation of simulation trajectories. We did 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 needed to identify and interpret the key factors behind the good performance of our "causal" classifiers. The results presented above led us to conclude that the most important factors appeared to be, first, external events acting through fear (F_E), followed by social influence acting through both anger (A_S) and fear (F_S). This was the kind of hint that machine learning could extract from our causal partitions. In order to proceed further, we needed to incorporate deeper technical knowledge of our agent-based model into our hypotheses.

Our initial hypothesis was that early in the scenario, as a result of some unknown process or random fluctuation, there formed a nucleus of agents that were either angry or afraid (while the other agents were still driven by their standard motives) and this nucleus then "converted" the rest of the agents by their social influence. If this was the case, we would expect the social influence components A_S and F_S to be negatively correlated (i.e., working against each other) and at the same time to be the best predictors for classification. Analysis of the simulation data, however, showed only a small (albeit negative) correlation. Moreover, this pair was not the best predictor, since its combined accuracy was only about 95%. At this point our method was not yet so mature as to enable us to reject this hypothesis outright, but its likelihood decreased. The main weakness of our approach was that we causally partitioned only the final values of fear and anger at the end of the simulation, while the really "decisive" period seemed to be its early part. We felt the need to dynamically identify the moment in which the trajectory bifurcation began and apply causal partitioning 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. As a rule they selected the closest soldier as their target, and if they hit him twice, they were in turn immobilized by an effective shot. Thus, in the worst case, they only threw two stones, while in the best case, four (with three soldiers, the fourth stone-throw always resulted in immobilization). Given that stone-throws incite anger and effective shots mainly fear, the proportion of stone-throws to effective shots in the early part of the scenario might be the tipping factor determining its subsequent aggressive or timid turn. If this hypothesis was true, then by adjusting the soldiers to use only warning shots we should make all the scenarios turn aggressive. We tested this experimentally, but the bifurcation 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 behaviour). The first counterargument had already been furnished by the initial clustering exercise, in which the system behaviour was shown to be robust, without undue sensitivity to the time step size. We expected high sensitivity if the observed bifurcation had been primarily due to random fluctuations. Furthermore, forcing our soldier agents to use only warning shots was a much more significant change and yet the bifurcation persisted. We could therefore safely conjecture that the bifurcation was caused by some stable and robust mechanism. This did not mean that the element of randomness played no role (in fact it had to because without it all the simulations would have yielded an identical result) but that there were likely other deterministic factors amplifying and stabilizing the bifurcation.

Our fourth hypothesis was that the external events and social influence acted together, perhaps as part of a two-stage or multistage process. However, in order to verify it, we needed to improve our method first. The improvements and the new results are described in the next subsection.

5.2. Method Improvements and New Results. The first improvement aimed at the identification of the decisive moment when the trajectory bifurcation began. We have solved this by logging causal partitions periodically every two seconds. Later, off-line, we then identified the point when the partition components started exhibiting increased predictive power.

The second improvement reflected our need for more detailed information. Instead of the combined effect of all external events lumped together, we recorded the effect of civilian actions separately from that of security actions by splitting each "external event" component into two: F_E was split into F_{EC} (the effect of civilian actions) and F_{ES} (the effect of security actions), and A_E likewise into A_{EC} (civilian) and A_{ES} (security). The causal partitions of anger and fear thus became

$$\begin{aligned} F &= \left(F_{F0}, F_{EC}, F_{ES}, F_S\right), \\ A &= \left(A_{A0}, A_{EC}, A_{ES}, A_S, A_L\right) \end{aligned} \tag{24}$$

in accordance with the theoretical derivation of our method in Section 4.2.2.

Our new set of experiments started with 350 simulation runs with the time step $\Delta t = 300$ ms. Each run again covered 90 seconds of simulated time and periodically logged the causal partitions as well as the other relevant attributes N_E, N_W, N_S, *A-count*, and *F-count*. After a preliminary review of simulation logs, we rejected four for showing signs of numerical instability. Out of the 346 remaining ones, 196 belonged to the aggressive cluster and 150 to the timid cluster. Thus, even the trivial classifier assigning all the simulations to the aggressive cluster could achieve approximately 56.7% accuracy (196/346). This means we should consider as informative only those causal partitions and classifiers that achieve higher accuracy than 56.7%.

The shape of the two clusters, shown in Figure 4, is similar to that reported in [8]: timid scenarios congregate in the top left corner, while the aggressive ones appear to protrude from the bottom right corner toward the center of the figure.

The periodic logging of causal partitions allowed us to examine the time-evolution of their predictive power. It is shown graphically in Figure 5 and tabulated in Table 4. The 10th second of simulated time turned out to be the earliest moment when the outcome at the 90th second could be predicted with an increased accuracy, mainly thanks to F_{EC} (the effect of civilian actions on fear). In the 12th second, the prediction accuracy further increased, but here the importance of F_{EC} faded, having been replaced by F_{ES} (the effect of security actions on fear). In the 14th and the 16th

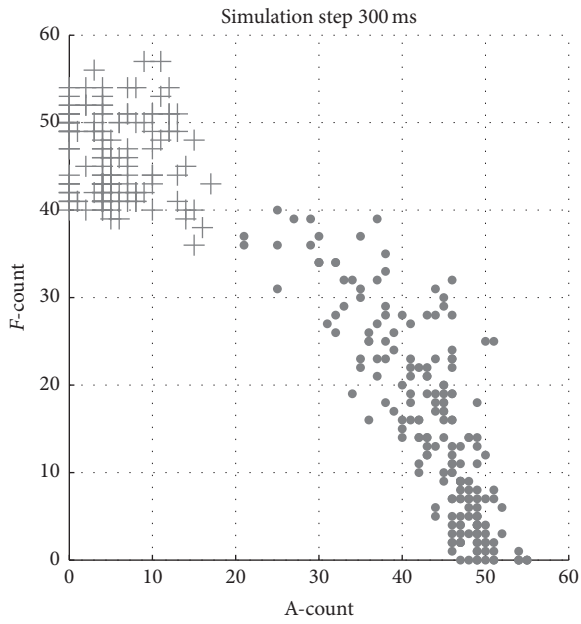

Figure 4: Simulation clusters for the new set of experiments.

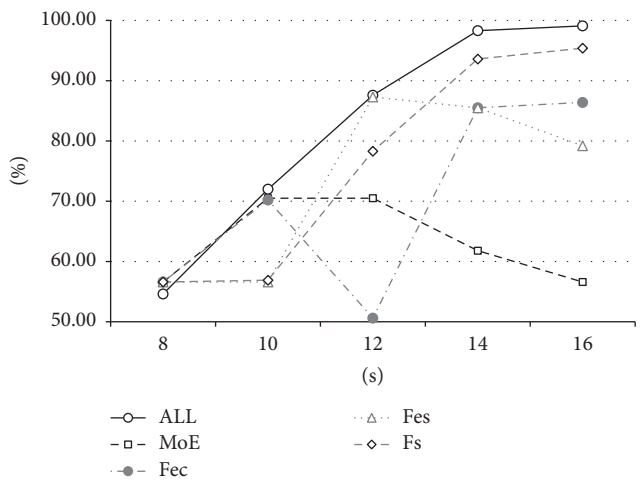

Figure 5: Time-evolution of the accuracy of selected predictors.

seconds, the prediction accuracy came close to 100%, but here the lead shifted to F_S (the effect of social influence on fear). Limitations of our graphical software forced us to label F_{EC} as "Fec," F_{ES} as "Fes," and F_S as "Fs" in the figure. The label "ALL" represents the predictor built from all the nine causal partition components, while "MoE" stands for the predictor built from the variables N_E, N_W, N_S.

Several interesting observations can be drawn from this figure and table:

(1) The accuracy of the "ALL" causal model rose very quickly toward the theoretical maximum of 100%. We considered this very significant and promising.

(2) The competing "MoE" model could not keep the pace: its accuracy got stuck at 70% and later even decreased. Of course, the decrease was only temporary: our earlier experiments mentioned in Section 5.1 had shown MoE to regain nearly 100% accuracy toward the end of the simulation.

(3) The "MoE" predictor was not the only one whose predictive power did not monotonously increase with time; our causal predictors F_{EC}, F_{ES} also fluctuated. Particularly in the case of F_{ES} we suspected that it was "polluted" with irrelevant information after the 12th second.

(4) The succession in which the individual partition components took the lead in predictive accuracy appeared to confirm our suspicion of a staged process. Its first phase seemed to be related to civilian actions; this would explain why F_{EC} took the lead at the 10th second. Its second phase appeared to be linked with security actions, conferring on them the predictive lead at the 12th second. Finally, social influence seemed to take over at the 14th second, conferring the lead on F_S.

Table 4: Time-evolution of the accuracy of selected predictors tabulated.

Predictor	Time				
	8 s	10 s	12 s	14 s	16 s
ALL	54.6	72.0	87.6	98.3	99.1
MoE	56.6	70.5	70.5	61.8	56.6
F_{EC}	56.6	70.2	50.6	85.5	86.4
F_{ES}	56.6	56.6	87.3	85.5	79.2
F_S	56.6	56.9	78.3	93.6	95.4

It is safe to say that the time-evolution of the predictive power of partition components and other relevant attributes, whether displayed graphically as in Figure 5 or tabulated as in Table 4, would be the principal tool of investigators using the method of causal partitions. The key to success is the ability to map the peaks in predictive accuracy to their potential causes or at least to suspected phases or stages of the observed emergence. In this case, our knowledge of this particular model helped us map two of the three suspected phases mentioned above to what we knew about the behaviour of our agents:

(1) It was always the aggressive civilians that started the confrontation by stone-pelting the approaching soldier patrol, so this might be the first phase.

(2) The patrol members simply reacted to stone-pelting by warning and effective shots, which might then be the second phase.

This sequence was common to all our simulations. Afterwards they bifurcated along two different routes. Regardless of the route taken, the process seemed to be linked with social influence, and this might then constitute a hypothetical third phase.

At this point we felt to be tantalizingly close to the solution, yet we could not quite put our finger on it. We were looking for the feature of our model that was causing the bifurcation, so that by adjusting it we could force all the simulations along one common route. Moreover, it had to be something that differentiated the "timid" simulations from the "aggressive" ones. We noticed the first signs of differentiation in the third phase linked with social influence and conjectured that the aggressive scenarios were fuelled by social influence on anger while the timid ones by social influence on fear. But even this appeared to be an ex post phenomenon, a consequence of some invisible factor that just kept eluding us.

As in our earlier experiments, we were ultimately forced to come up with hypotheses incorporating deeper technical knowledge of our model. But this time we had one more precious clue at our disposal: whatever the identity of the elusive factor, it was clearly acting between the 8th second and the 14th second of simulated time. We focused on this early period and, using the graphs of selected model variables provided by the MASON simulation framework, we managed to identify a new "prime suspect": by the 14th second, sensory perception of most of our agents was already blocked. Although significant, it was again something common to all our simulations and, by itself, could not differentiate the "timid" from the "aggressive" ones. But in conjunction with other known facts it finally helped us to piece together a more adequate picture of the mechanism of emergence. Our new hypothesis postulated that when the aggressive civilians start stone-pelting the soldiers, the fate of the scenario depends on how quickly and resolutely the soldiers react *before* the sensory perception of the civilians gets blocked. Later actions will not affect their emotions. There are two principal sources of variability in the scenario which impact the initial attack and the subsequent security response:

(1) We slightly randomize the duration of the agents' actions so as to prevent massive synchronised replanning of their behaviour which we found to interfere with smooth visual rendering and real-time operation of our system.

(2) At the beginning of each simulation, civilian agents are created in random positions near the shop. Before they can attack the soldiers, the aggressive civilians have to find shelter behind the corners of the buildings. Due to their random initial positions, the exact timing and intensity of their attack are slightly different in each simulation run.

It might then be the variable sequence of events perceived by the civilians before their perception gets blocked that seals the fate of each simulation. In order to illustrate this principle, let us consider a couple of event sequences and their emotional impact as per Table 2. At one extreme we have six successive stone-throws: these are enough to block the perception of the civilians and incite in them a lot of anger but very little fear. Thus, whenever the aggressive individuals manage to mount a quick and intensive assault, such scenarios will tend to turn aggressive. The opposite extreme is not so clear-cut because the soldiers cannot start firing without a cause. The confrontation has to start with a stone-throw, followed by a warning shot, followed by another stone-throw, followed by an effective shot. This event sequence too blocks the perception of the civilians and incites in them the maximum achievable amount of fear. Consequently, such scenarios should be "timid."

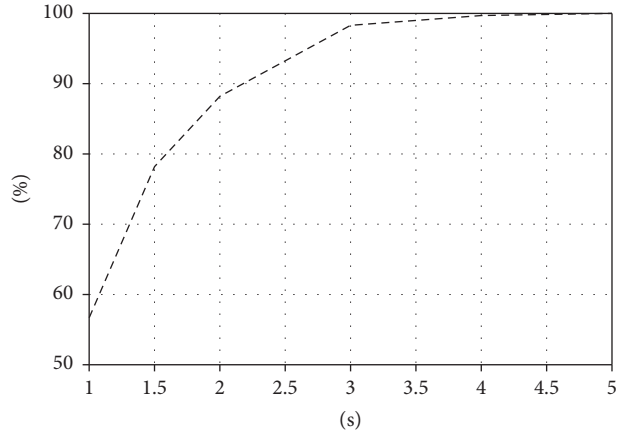

Figure 6: Proportion of "aggressive" simulations as a function of soldier reaction time.

If this hypothesis was correct, then we should be able to control the fate of our simulations by adjusting the reaction time of soldier agents. More specifically, by prolonging their reaction time, we should exacerbate the civilian tendency toward aggression. Soldier reaction time is a parameter which is not obviously linked with either fear or anger and, had we not investigated the process of emergence, we would have had little reason to suspect that it played a significant role in it.

In order to verify this hypothesis, we run several batches of simulations, each consisting of 350 simulation runs and each with a different setting of soldier reaction time. The results, shown in Figure 6, confirmed our hypothesis. The standard value of soldier reaction time was one second, at which 56.7% of the simulation runs in the batch turned aggressive. Increasing this time by just half a second increased the proportion of aggressive simulations to 78.2%. At two seconds, the proportion was already 88.2%, at three seconds 98.3%, and at four seconds 99.7%. At five seconds, all the simulations turned aggressive.

We would like to point out here that we were not surprised simply by the fact that soldier reaction time affected the subsequent development of simulation scenarios: that was only natural and expected. What surprised us was that the reasons for the observed bifurcation fell into this category too. This at first appeared completely counterintuitive since in both timid and aggressive simulations the soldiers reacted equally quickly when measured in absolute time; for example, in our baseline simulation batches (with soldier reaction time set to one second), it was precisely one second after being hit by a stone that soldier agents emitted either a warning or an effective shot, so how could this be at once "too late" for simulations that turned aggressive and swift enough for those that turned timid? Our surprising realisation lay in

the fact that in this case of emergence the speed of response was not to be judged by the time it actually took but rather relatively to stones thrown, and the blocking of the sensory perception of civilian agents was the key "accessory" that helped shape the system response in this peculiar way. A somewhat simplified expression of this dependence might be the following: "How many stones were thrown by the time the first warning shot occurred?" If too many, then the simulation would turn aggressive.

Although we could in principle formulate further experiments to elucidate finer aspects of the behaviour of our model, our main task is now essentially completed. Regarding the bifurcation of simulation trajectories, we have uncovered its phases as well as its generating process and have even identified a nonobvious model feature through which we could suppress it.

6. Conclusion and Future Work

In this article, we introduced the concept of a *causal partition* of a model variable and developed the method of causal partitioning which can be used to investigate emergent phenomena in complex systems.

We derived the method in the context of an agent-based model of human behaviour from project EUSAS, in which simulated emotions of *fear* and *anger* competed for the control of agent behaviour and were partly driven by ordinary nonlinear differential equations. In one of our scenarios, this model exhibited a puzzling emergent behaviour: simulations with the same input parameter setting bifurcated along two different trajectories. The time-evolution of the predictive power of causal partitions with respect to this bifurcation helped us identify its stages as well as its generating process. This in turn helped us find a nonobvious model feature through which we could suppress the bifurcation and force all the simulations along one common route.

In general, the method of causal partitions comprises four steps. The first, mathematically rigorous, includes causal analysis and the derivation of causal partitions for the system of interest. As a result, causal partitions produced by simulations in the second step can be reliably interpreted as showing how much each causal factor contributed to the numerical value of the partitioned model variables. The third and the fourth steps are more heuristic in character and combine elements of statistics, data mining, and machine learning. This phase starts with the investigation of the predictive accuracy of causal partitions with respect to the observed emergent phenomenon. It is important to keep in mind that high predictive accuracy of a given partition component does *not* by itself guarantee that the causal factor behind it is actually causing the phenomenon: it may be merely associated with it. Additional validation activities are needed, separate experiments manipulating the suspected cause and confirming or disconfirming its effect on the emergent phenomenon. In any case, the best early predictors help us focus on the relevant aspects of the system in search for its real causes and generating processes.

It will also be well to point out that our approach can be applied only to systems modelled through structural equations. This constrains the type of agent-based systems and the types of emergence that can be investigated with it. We do not claim that our approach can throw light on all kinds of emergence in agent-based systems. Specifically, we at present view the emergent behaviour of simple rule-based agents as outside its scope: in order to apply causal analysis to this case we would have to define, first, what we mean by cause and effect in such a system and, second, how we quantify causal effects.

Regarding future work, we see several directions. The first is to investigate our agent-based model in greater depth and elucidate certain subsidiary aspects before we apply our method to other systems. The first such aspect is the significance of social influence for the bifurcation of simulation trajectories. At present, we tend to think that social influence is not crucial; that is, the bifurcation would persist even if we kept I_F and I_A in (1) and (2) at zero level, for example, by setting the constant c_{4V} in (3) to zero. Furthermore, we do not expect that this would significantly affect the time-evolution of the predictive power of partition components shown in Figure 5. What we expect to change is the composition of the clusters; that is, some simulations might shift their cluster affiliation, although we are at present unable to say which cluster would grow and which would shrink. We also expect this effect to be relatively mild (if any). Another interesting question is how many clusters are really there. In one of our early reports [21] we decided to work with two clusters, but that was an arbitrary decision in order to keep the analysis simple. We were not at all sure that we would succeed and were ready to consider more clusters if necessary. The fact that we *could* complete the analysis using only two clusters does not by itself settle the question; it might merely show that our method tolerates some uncertainty in this respect.

The second direction, and a natural next step, is to apply our method to other similar systems. By similarity we mean that, besides being modelled through structural equations, the variables and functions used in them should be numerical (ideally, real-valued). The most straightforward application would be to systems governed by ordinary differential equations, but we believe an extension to systems driven by partial differential equations should be possible as well. As already mentioned, agent-based models using structural equations would make another legitimate and interesting analytical target. We also see an exciting possibility to adapt our method for use in artificial neural networks: these too are described by structural equations, because each neuron has only one output (dependent variable) which is a function of one or more inputs. In this way we might be able to study, for example, the processes of learning in complex neural architectures.

The discipline of System Dynamics merits special mention: it not only deals with dynamical systems but also painstakingly maps the flows of causality through them. Therefore, in most cases, SD model's equations should qualify as structural equations, which would make causal partitioning applicable to them at least in principle. As an example, let us consider variable X (modelled in SD as a *stock*) with n inflows and outflows $x_1, x_2, \ldots, x_n$. For simplicity, let us assume that the dynamical process we are interested in starts

at time $t = 0$ with zero initial value of X. Then its value at an arbitrary subsequent point of time t can be expressed as the sum of all its inflows and outflows separately integrated (or summed) over the time interval $[0, t]$. Thus, if X_i denotes the integrated or summed flow x_i over the concerned period, then $X(t) = X_1 + \cdots + X_n$. We can then represent $X(t)$ by a vector-like structure $(X_1, \ldots, X_n)$ which closely corresponds to what we have called a *basic causal partition* earlier in the article. In this "basic" version of causal partitioning, we look on each flow as contributing to its own dedicated partition component of X. The purpose of partitioning is to gain additional information: we now know not only the resulting value of X but also how much each flow contributed, which might help us understand some of the more puzzling behaviours of the modelled system, for example, through exploring the time-evolution of the predictive power of partition components and their combinations with respect to the observed behaviour. More sophisticated ("enhanced") forms of causal partitioning might try to redefine causal partitions typically by splitting some components into two or more in order to gain more information or by eliminating others in order to improve interpretability.

Finally, the last direction of future work comprises the study of the method's mathematical properties, especially the limits of its stability. Elsewhere in the article we mentioned that we had to reject some data because it exhibited signs of numerical instability. We could not go into details, but in [22] we mentioned how this could be handled, and we intend to follow it up alongside our work on further practical applications.

Appendix

A. Applicability of Structural Causality to Ordinary Differential Equations

In this appendix we restrict our attention to (1)-(2) and their generalised form ((4a) and (4b)). Following the notation used in the "prototypical" structural equation (9), such equations can be written as

$$\frac{dY_k}{dt} = g_k\left(\widetilde{pa}_k, u_k\right), \quad k = 1, \ldots, n, \tag{A.1}$$

where the function g_k can be nonlinear. One potential problem seems to be that instead of a model variable there is a time derivative on the left-hand side; another problem is that the dependent variable Y_k typically influences its own time derivative; that is, it appears to belong to its own "parent set": $Y_k \in \widetilde{pa}_k$. In order for such models to qualify as "structural," the differential equations need to be converted into difference equations, for example, by approximating the derivatives by difference quotients:

$$\frac{dY_k}{dt} \approx \frac{\Delta Y_k^{(j)}}{\Delta t} = \frac{Y_k^{(j)} - Y_k^{(j-1)}}{t_j - t_{j-1}}, \quad k = 1, \ldots, n, \; j = 1, \ldots, m, \tag{A.2}$$

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. Substituting from (A.2) into (A.1) leads to

$$Y_k^{(j)} \approx Y_k^{(j-1)} + \Delta t \cdot g_k\left(\widetilde{pa}_k^{(j-1)}, u_k^{(j-1)}\right), \quad k = 1, \ldots, n, \; j = 1, \ldots, m. \tag{A.3}$$

Now the relationship to (9) becomes clearer. The left-hand side of (A.3) corresponds to variable X_k in (9), which means that in this "structural form" the value of Y_k at time t_j is considered a separate "structural" variable, distinct from the values of Y_k at other points in time. Most importantly, in this form, the parent set of $Y_k^{(j)}$ no longer contains this variable but only the preceding values of Y_k in time. Thus, after the discretization, no "structural" variable depends on itself. We can therefore conclude that the discretized version of our model does qualify as a structural model as defined by Pearl in [5].

B. Improving and Fine-Tuning Causal Partitions

Regarding mathematical notation used in this appendix, we view the process of simulation as effectively discretizing time into a sequence $t_0, t_1, t_2, \ldots$, whose general member t_j can be calculated as $t_j = j \cdot \Delta t$. The first simulation step then covers the period $[t_0, t_1)$ or $[0, \Delta t)$. In general, the jth simulation step spans the period $[t_{j-1}, t_j)$ which can be calculated as $[(j-1) \cdot \Delta t, j \cdot \Delta t)$. With respect to model variables we employ superscript indexing where $X^{(j)}$ stands for $X(t_j)$. Analogously, $\Delta X^{(j)}$ denotes its backward difference at time t_j: $\Delta X^{(j)} = X^{(j)} - X^{(j-1)}$.

B.1. Eliminating a Redundant Partition Component. We show here how we managed to eliminate the component F_F of the basic causal partition of fear defined in Section 4.1. The elimination rests on a revised version of (15), in which the term ΔF has been further analysed. Practically, ΔF represents the change of F over one simulation step which spans a period Δt of simulated time.

If we apply (15) to backward differences at time t_{j-1} and expand the term $\Delta f^{(j-1)}$, with a slight rearrangement, we get

$$f^{(j-1)} \approx f^{(j-2)} + \Delta F^{(j-1)} \cdot \left.\frac{\partial f}{\partial F}\right|_{\substack{F=\widehat{F}^{(j-1)} \\ I_F=\widehat{I}_F^{(j-1)}}} + \Delta I_F^{(j-1)} \cdot \left.\frac{\partial f}{\partial I_F}\right|_{\substack{F=\widehat{F}^{(j-1)} \\ I_F=\widehat{I}_F^{(j-1)}}}. \tag{B.1}$$

Here, $f^{(j-1)}$ is the value of f at the beginning of the current (jth) simulation step, and $f^{(j-2)}$ is its value at the beginning of the previous step. Partial derivatives are evaluated at midpoint $(\widehat{F}^{(j-1)}, \widehat{I}_F^{(j-1)})$ of the previous step. Our goal is to causally partition the total value of fear at the end of the current simulation step $F_T^{(j)}$. Equation (B.1) can be interpreted as saying that the first derivative of fear at the

beginning of the current step can be approximated on the basis of the values from the previous step. Of course, we do not need (B.1) in order to calculate $f^{(j-1)}$ because, knowing $F_T^{(j-1)}$ and $I_F^{(j-1)}$, we can calculate it directly. The reason we need (B.1) is to causally partition f, which is a prerequisite for causally partitioning fear itself.

At the same time, $\Delta F^{(j-1)}$ is simply the sum of the discrete and the continuous changes during the previous simulation step:

$$\Delta F^{(j-1)} = c_{5F} \cdot \Delta E_F^{(j-1)} + f^{(j-2)} \cdot \Delta t. \tag{B.2}$$

Substituting this into (B.1) eliminates fear F as a cause of its first derivative f:

$$\begin{aligned} f^{(j-1)} \approx\ & f^{(j-2)} \\ & + \left(c_{5F} \cdot \Delta E_F^{(j-1)} + f^{(j-2)} \cdot \Delta t\right) \cdot \left.\frac{\partial f}{\partial F}\right|_{\substack{F=\widehat{F}^{(j-1)} \\ I_F=\widehat{I}_F^{(j-1)}}} \\ & + \Delta I_F^{(j-1)} \cdot \left.\frac{\partial f}{\partial I_F}\right|_{\substack{F=\widehat{F}^{(j-1)} \\ I_F=\widehat{I}_F^{(j-1)}}}. \end{aligned} \tag{B.3}$$

This can be rewritten as

$$f^{(j-1)} \approx c_{1S} \cdot f^{(j-2)} + c_{2S} \cdot c_{5F} \cdot \Delta E_F^{(j-1)} + c_{3S} \cdot \Delta I_F^{(j-1)}, \tag{B.4}$$

where c_{1S}, c_{2S}, c_{3S} represent sensitivities or weighting factors based on the partial derivatives evaluated at the midpoint $(\widehat{F}^{(j-1)}, \widehat{I}_F^{(j-1)})$ of the previous simulation step:

$$c_{3S} = \left.\frac{\partial f}{\partial I_F}\right|_{\substack{F=\widehat{F}^{(j-1)} \\ I_F=\widehat{I}_F^{(j-1)}}}, \tag{B.5a}$$

$$c_{2S} = \left.\frac{\partial f}{\partial F}\right|_{\substack{F=\widehat{F}^{(j-1)} \\ I_F=\widehat{I}_F^{(j-1)}}}, \tag{B.5b}$$

$$c_{1S} = 1 + c_{2S} \cdot \Delta t. \tag{B.5c}$$

These partial derivatives can be approximated numerically.

Equation (B.4) can be directly translated into an improved method of causal partitioning. In this version of the method, the first derivative of fear f is represented by the causal partition vector (f_E, f_S) and fear F by the partition (F_E, F_S). Partition components indexed by "E" represent the contribution of external events, while those indexed by "S" stand for the contribution of social influence.

The improved method proceeds as follows: in order to causally partition the first derivative of fear at the beginning of the current simulation step $f^{(j-1)}$, take its previous partition $f^{(j-2)}$, multiply it by c_{1S}, and add the weighted contributions of external events and social influence during the *previous* simulation step to their respective partition components f_E and f_S. This can be expressed through recurrence relations, one for each partition component:

$$f_E^{(j-1)} = c_{1S} \cdot f_E^{(j-2)} + c_{2S} \cdot c_{5F} \cdot \Delta E_F^{(j-1)}, \tag{B.6a}$$

$$f_S^{(j-1)} = c_{1S} \cdot f_S^{(j-2)} + c_{3S} \cdot \Delta I_F^{(j-1)}. \tag{B.6b}$$

In practice, we do not even need (B.6b) since we can handily calculate f_S from the causal partition constraint: $f_S = f - f_E$.

Having partitioned the first derivative of fear f, we can now partition fear itself. In order to partition its new total value at the end of the current simulation step $F_T^{(j)}$, take its previous partition $F_T^{(j-1)}$ and add to it, component by component, the "continuous" increment for the current simulation step. This increment can be written in vector form as $(\Delta t \cdot f_E^{(j-1)}, \Delta t \cdot f_S^{(j-1)})$. Then add the cumulative impact of external events perceived during the current step $c_{5F} \cdot \Delta E_F^{(j)}$ directly to F_E. Expressed in the form of recurrence relations it becomes

$$F_E^{(j)} = F_E^{(j-1)} + \Delta t \cdot f_E^{(j-1)} + c_{5F} \cdot \Delta E_F^{(j)}, \tag{B.7a}$$

$$F_S^{(j)} = F_S^{(j-1)} + \Delta t \cdot f_S^{(j-1)}. \tag{B.7b}$$

Again, instead of calculating F_S as per (B.7b), we can calculate it more handily from the causal partition constraint: $F_S = F_T - F_E$.

B.2. Handling Dependency among Causal Factors. On the basis of (21), we can approximate the total effect of fear F on the value of f as the sum of its direct effect and its indirect effect:

$$\frac{df}{dF} \cdot \Delta F = \underbrace{\frac{\partial f}{\partial F} \cdot \Delta F}_{\text{direct}} + \underbrace{\frac{\partial f}{\partial I_F} \frac{dI_F}{dF} \cdot \Delta F}_{\text{indirect}}. \tag{B.8}$$

We now need to restructure (15) in line with this more general perspective, which can be done by simultaneously adding and subtracting the indirect effect of fear on its right-hand side:

$$\begin{aligned} \Delta f \approx\ & \underbrace{\left[\frac{\partial f}{\partial F} \cdot \Delta F + \frac{\partial f}{\partial I_F} \frac{dI_F}{dF} \cdot \Delta F\right]}_{\text{total effect of fear}} \\ & + \underbrace{\left[\frac{\partial f}{\partial I_F} \cdot \Delta I_F - \frac{\partial f}{\partial I_F} \frac{dI_F}{dF} \cdot \Delta F\right]}_{\text{social influence residual}}. \end{aligned} \tag{B.9}$$

We see that Δf, which represents the total change of f over one simulation step, can be approximated (and therefore interpreted) as the total effect of fear F of our agent plus a term which we propose to call the *social influence residual* and denote by Δf_{SR}. We obtain this residual by adjusting the "gross" effect of social influence for the indirect effect of fear F, so that it only includes the effect of fear of the remaining (influencing) agents. If we express the total effect of F through the total derivative and compact the residual, we obtain

$$\Delta f \approx \frac{df}{dF} \cdot \Delta F + \Delta f_{SR}. \tag{B.10}$$

This equation holds for all simulation steps. As in the derivation of the enhanced version of the method in Section B.1, we apply it to the simulation step *preceding* the

current one; that is, we use backward differences at time t_{j-1}. Expanding $\Delta f^{(j-1)}$ and substituting for $\Delta F^{(j-1)}$ from (B.2) lead to a generalised form of (B.4) which does not require F and I_F to be independent:

$$\begin{aligned} f^{(j-1)} &\approx f^{(j-2)} \\ &\quad + \left(c_{5F} \cdot \Delta E_F^{(j-1)} + f^{(j-2)} \cdot \Delta t\right) \cdot \left.\frac{df}{dF}\right|_{\substack{F=\widehat{F}^{(j-1)} \\ I_F=\widehat{I}_F^{(j-1)}}} \qquad \text{(B.11a)} \\ &\quad + \Delta f_{SR}^{(j-1)} \\ &\approx c'_{1S} \cdot f^{(j-2)} + c'_{2S} \cdot c_{5F} \cdot \Delta E_F^{(j-1)} + \Delta f_{SR}^{(j-1)}, \qquad \text{(B.11b)} \end{aligned}$$

where c'_{1S}, c'_{2S} are sensitivities or weighting factors capturing the *total* sensitivity of f to changes in F (as opposed to the *direct* sensitivity captured by c_{1S}, c_{2S} in (B.5a), (B.5b), and (B.5c)):

$$c'_{2S} = \left.\frac{df}{dF}\right|_{\substack{F=\widehat{F}^{(j-1)} \\ I_F=\widehat{I}_F^{(j-1)}}}, \qquad \text{(B.12a)}$$

$$c'_{1S} = 1 + c'_{2S} \cdot \Delta t. \qquad \text{(B.12b)}$$

This leads to recurrence relations for components of f resembling those of (B.6a) and (B.6b):

$$f_E^{(j-1)} = c'_{1S} \cdot f_E^{(j-2)} + c'_{2S} \cdot c_{5F} \cdot \Delta E_F^{(j-1)}, \qquad \text{(B.13a)}$$

$$f_S^{(j-1)} = c'_{1S} \cdot f_S^{(j-2)} + \Delta f_{SR}^{(j-1)}. \qquad \text{(B.13b)}$$

As a result of working with *total* sensitivities, the component f_S now represents just the social influence residual, that is, only the effect of fear of the influencing agents.

Regarding fear F, its recurrence relations remain the same as in (B.7a) and (B.7b), and the partition constraint $F_T = F_E + F_S$ continues to hold as well.

B.3. Splitting Partition Components to Gain More Information. In this section, we illustrate the process of splitting partition components on the component F_E which in the previous versions of causal partitioning held the total contribution of all the perceived events regardless of their type. If we sum the effects of perceived events separately by type, we can express their total effect as

$$\Delta E_F = \Delta E_{F0} + \Delta E_{FC} + \Delta E_{FS}, \qquad \text{(B.14)}$$

where ΔE_{F0} is the effect of the setting of the initial value of fear F_{F0}, ΔE_{FC} is the effect of civilian actions, and ΔE_{FS} is that of security countermeasures. Substituting this into (B.11b) provides justification for further splits (or fine-tuning) of our causal partitions:

$$\begin{aligned} f^{(j-1)} \approx c'_{1S} \cdot f^{(j-2)} + c'_{2S} \cdot c_{5F} \\ \cdot \left(\Delta E_{F0}^{(j-1)} + \Delta E_{FC}^{(j-1)} + \Delta E_{FS}^{(j-1)}\right) + \Delta f_{SR}^{(j-1)}. \end{aligned} \qquad \text{(B.15)}$$

The above equation enables us to split the original causal partition component f_E as defined in Section B.2 into three: f_{F0}, f_{EC}, f_{ES}. The idea of "splitting" means that the recurrence relations for these new derived components will be analogous to that for f_E in (B.13a), except that each new component now only includes the appropriate kind of event impact:

$$f_{F0}^{(j-1)} = c'_{1S} \cdot f_{F0}^{(j-2)} + c'_{2S} \cdot c_{5F} \cdot \Delta E_{F0}^{(j-1)}, \qquad \text{(B.16a)}$$

$$f_{EC}^{(j-1)} = c'_{1S} \cdot f_{EC}^{(j-2)} + c'_{2S} \cdot c_{5F} \cdot \Delta E_{FC}^{(j-1)}, \qquad \text{(B.16b)}$$

$$f_{ES}^{(j-1)} = c'_{1S} \cdot f_{ES}^{(j-2)} + c'_{2S} \cdot c_{5F} \cdot \Delta E_{FS}^{(j-1)}, \qquad \text{(B.16c)}$$

$$f_S^{(j-1)} = c'_{1S} \cdot f_S^{(j-2)} + \Delta f_{SR}^{(j-1)}. \qquad \text{(B.16d)}$$

Likewise, F_E will be split into F_{F0}, F_{EC}, F_{ES}, with recurrence relations analogous to that for F_E in (B.7a):

$$F_{F0}^{(j)} = F_{F0}^{(j-1)} + \Delta t \cdot f_{F0}^{(j-1)} + c_{5F} \cdot \Delta E_{F0}^{(j)}, \qquad \text{(B.17a)}$$

$$F_{EC}^{(j)} = F_{EC}^{(j-1)} + \Delta t \cdot f_{EC}^{(j-1)} + c_{5F} \cdot \Delta E_{FC}^{(j)}, \qquad \text{(B.17b)}$$

$$F_{ES}^{(j)} = F_{ES}^{(j-1)} + \Delta t \cdot f_{ES}^{(j-1)} + c_{5F} \cdot \Delta E_{FS}^{(j)}, \qquad \text{(B.17c)}$$

$$F_S^{(j)} = F_S^{(j-1)} + \Delta t \cdot f_S^{(j-1)}. \qquad \text{(B.17d)}$$

Finally, we also need to adjust the causal partition constraints:

$$f^{(j)} = f_{F0}^{(j)} + f_{EC}^{(j)} + f_{ES}^{(j)} + f_S^{(j)}, \qquad \text{(B.18a)}$$

$$F_T^{(j)} = F_{F0}^{(j)} + F_{EC}^{(j)} + F_{ES}^{(j)} + F_S^{(j)}. \qquad \text{(B.18b)}$$

These more detailed causal partitions enable us to examine the effect of civilian actions on the emergent behaviour of the system separately from that of security countermeasures or the setting of the initial value of fear.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by EDA Project A-0938-RT-GC EUSAS, national project VEGA no. 2/0167/16, and the Slovak Research and Development Agency under the Contracts nos. APVV-0233-10 and APVV-0809-11. The authors would like to thank Dr. Ladislav Halada of the Institute of Informatics of the Slovak Academy of Sciences for helpful suggestions regarding mathematical notation.

References

[1] R. G. Sargent, "Verification and validation of simulation models," *Journal of Simulation*, vol. 7, no. 1, pp. 12–24, 2013.

[2] M. A. Louie and K. M. Carley, "Balancing the criticisms: validating multi-agent models of social systems," *Simulation Modelling Practice and Theory*, vol. 16, no. 2, pp. 242–256, 2008.

[3] M. Luck, P. McBurney, O. Shehory, and S. Willmott, "Agent technology roadmap: a roadmap for agent based computing," Tech. Rep., University of Southampton on behalf of AgentLink III, 2005, http://eprints.soton.ac.uk/261788/1/al3roadmap.pdf.

[4] S. Bullock and D. Cliff, "Complexity and emergent behaviour in ICT systems," Tech. Rep., Hewlett-Packard Labs, Palo Alto, Calif, USA, 2004, http://eprints.soton.ac.uk/261478/1/HPL-2004-187.pdf.

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

[6] J. Y. Halpern and J. Pearl, "Causes and explanations: a structural-model approach. part I: causes," *British Journal for the Philosophy of Science*, vol. 56, no. 4, pp. 843–887, 2005.

[7] J. Y. Halpern and J. Pearl, "Causes and explanations: a structural-model approach. Part II: explanations," *British Journal for the Philosophy of Science*, vol. 56, no. 4, pp. 889–911, 2005.

[8] M. Kvassay, L. Hluchý, P. Krammer, and B. Schneider, "Causal analysis of the emergent behavior of a hybrid dynamical system," *Acta Polytechnica Hungarica*, vol. 11, no. 4, pp. 21–40, 2014.

[9] M. Kvassay, L. Hluchý, Š. Dlugolinský et al., "A novel way of using simulations to support urban security operations," *Computing and Informatics*, vol. 34, no. 6, pp. 1201–1233, 2015.

[10] L. Berkowitz, Ed., *Aggression: A Psychological Analysis*, McGrawHill, New York, NY, USA, 1962.

[11] S. Prentice-Dunn and R. W. Rogers, "Deindividuation and the self regulation of behaviour," in *Psychology of Group Influence*, P. B. Paulus, Ed., pp. 89–109, Lawrence Erlbaum Associates, Hillsdale, Mich, USA, 1989.

[12] E. Staub, "Predicting collective violence: the psychological and cultural roots of turning against others," in *Collective Violence*, C. Summers and E. Markusen, Eds., pp. 195–209, Rowman & Littlefield Publishers, Lanham, Md, USA, 1999.

[13] D. Cañamero, "Modeling motivations and emotions as a basis for intelligent behaviour," in *Proceedings of the 1st International Conference on Autonomous Agents (AGENTS '97)*, pp. 148–155, ACM Press, Marina del Rey, Calif, USA, February 1997.

[14] B. Latané, "Dynamic social impact," *Journal of Communication*, vol. 46, no. 4, pp. 13–25, 1996.

[15] 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., pp. 83–114, Physica Verlag, 2000.

[16] B. Schmidt, "The modelling of human behaviour: the PECS reference model," in *Proceedings of the 14th European Simulation Symposium and Exhibition (ESS '02)*, A. Verbraeck and W. Krug, Eds., pp. 13–18, SCS Europe BVBA, Dresden, Germany, October 2002.

[17] C. Swinerd and K. R. McNaught, "Design classes for hybrid simulations involving agent-based and system dynamics models," *Simulation Modelling Practice and Theory*, vol. 25, pp. 118–133, 2012.

[18] J. Pearl, "An introduction to causal inference," *International Journal of Biostatistics*, vol. 6, no. 2, Art. 7, 61 pages, 2010.

[19] M. Kvassay, L. Hluchy, B. Schneider, and H. Bracker, "Towards causal analysis of data from human behaviour simulations," in *Proceedings of the IEEE International Symposium on Logistics and Industrial Informatics (LINDI '12)*, Smolenice, Slovakia, September 2012.

[20] Weka 3: Data mining software in Java, http://www.cs.waikato.ac.nz/ml/weka.

[21] M. Kvassay, L. Hluchy, P. Krammer, and B. Schneider, "Exploring human behaviour models through causal summaries and machine learning," in *Proceedings of the 17th IEEE International Conference on Intelligent Engineering Systems (INES '13)*, pp. 231–236, San Jose, Calif, USA, June 2013.

[22] M. Kvassay, L. Hluchý, and B. Schneider, "Summarizing the behaviour of complex dynamic systems," in *Proceedings of the IEEE 11th International Symposium on Applied Machine Intelligence and Informatics (SAMI '13)*, pp. 15–20, IEEE, Herľany, Slovakia, February 2013.