TU Delft

Delft University of Technology

Document Version
Final published version

Licence
Dutch Copyright Act (Article 25fa)

Citation (APA)
Dert, T., Balakrishnan, S., van der Wal, N., & Comes, T. (2026). What to Automate, When, and Why? A concept design to explore Human–AI Teaming in Crisis Management. *Proceedings of the International ISCRAM Conference*, *23*. https://doi.org/10.59297/57cnkh29

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

What to Automate, When, and Why? A concept design to explore Human–AI Teaming in Crisis Management

Tamara Dert
Delft University of Technology
t.l.dert@tudelft.nl

Srijith Balakrishnan
Delft University of Technology
s.balakrishnan@tudelft.nl

Natalie van der Wal
Delft University of Technology
c.n.vaderwal@tudelft.nl

Tina Comes
Delft University of Technology
t.comes@tudelft.nl

ABSTRACT

Artificial intelligence promises rapid information processing, analysis and decisions. Yet, guidance on what to automate, when, and to what degree, remains limited for Human-AI teams. Case-based empirical studies provide rich context, but a framework for systematic exploration of Human–AI team performance is missing. This paper introduces a conceptual model for Human–AI teaming that integrates levels of automation, trust dynamics, and organizational functions within a social networked, agent based perspective. Building on the crisis information management cycle, it models sensing, analysis, sharing, and decision-making as an iterative loop in which automation shapes latency, reliability, and trust. As a proof of concept, we developed a minimal model with results showing how automation regimes, forecast horizons, and trust configurations affect performance through the concept. The model provides a starting point for users to explore cascading effects, authority shifts, and trade-offs between performance and meaningful human control.

Keywords

Human-AI Teaming; Levels of Automation; Trust Dynamics; Agent-Based Modelling; Crisis Management;

INTRODUCTION

Artificial intelligence (AI) is increasingly used in crisis management, where decision-makers face high stakes, extreme uncertainty, time pressure, and rapidly evolving information environments (Comes 2024). Advances in sensing, predictive analytics, and automated decision support promise to accelerate crisis management and reduce the cognitive burden on humans (Lauras and Comes 2015; Dubey et al. 2022; Kordi and Ertz 2025). Recognising that AI systems are intertwined with organizational structures and social processes rather than operating in isolation (Behl et al. 2025), the growing automation raises foundational questions about *what* to automate, *when*, and *to what degree*?

Empirical case studies of AI in crisis and operational settings are insightful but fragmented across domains, technologies, and organizational contexts (Lythreatis et al. 2025). Research frequently targets specific tools or user groups, limiting comparability and external validity, since the performance of AI systems depends not only on technical accuracy but also on how humans interpret, trust, and coordinate around algorithmic outputs (Carter-Browne et al. 2021). Moreover, crises are marked by information scarcity, nonlinear escalation, and tightly coupled organizational dynamics, where small perturbations can cascade into coordination failures (Helbing and Mukerji 2012; Hempel et al. 2018) — dynamics that are hard to explore in bounded experimental settings. Addressing this is particularly important, as research on AI in humanitarian supply chains cautions that poorly governed automation can amplify inequity, bias, and accountability gaps (Behl et al. 2025).

Current research often emphasizes task-level efficiency rather than team- and system-level processes. Levels-of-automation frameworks offer useful taxonomies (Endsley 2017), but provide limited guidance for managing dynamic couplings and feedback loops within the social and organizational context (Jamieson and Skraaning 2017). Human–AI teaming studies, examining how humans and AI interact in pursuit of shared goals, point to differences in sensemaking and coordination between humans and AI (Carter-Browne et al. 2021), and to moderating influences such as experience, task division, and explainability (Vaccaro et al. 2024; Langer et al. 2021). Yet these insights remain only weakly connected to system-level crisis research on fragmentation, coordination breakdowns, and information bottlenecks (Schakel and Wolbers 2019; Wolbers et al. 2017; Van De Walle and Comes 2015).

This paper makes headway in addressing the lack of automation research at team- and system level by proposing a concept design and model to systematically explore Human–AI Teaming for the crisis information management cycle at system-level. Therefore, automation in this paper does not refer to a specific algorithmic architecture or AI system; rather, we distinguish among varying functionalities and degrees of autonomy that such systems may assume. The concept integrates organizational functions, trust dynamics, and automation levels within a social-networked, agent-based modelling approach. Rather than prescribing an optimal allocation of automation, the goal is to support systematic *what-if* exploration of risks-performance trade-offs across automation degrees, trust conditions, and coordination structures. We illustrate this approach with a minimal agent-based model that examines cascading effects and the emergence (or erosion) of trust under alternative design automation regimes in a crowd management setting.

BACKGROUND

AI for Crisis Information Management

Crisis information management research consistently shows that coordination problems stem less from information scarcity than from how information becomes fragmented, filtered, or unevenly shared across organizations Comes et al. (2020) and Van De Walle and Comes (2015). Within these socio-technical constraints, AI is positioned to accelerate detection, classification, sharing, and fusion across heterogeneous sources and users to enhance crisis management performance (Lauras and Comes 2015; Kordi and Ertz 2025). Empirical work in humanitarian logistics suggests that AI-driven analytics already contribute to agility, resilience, and overall performance (Dubey et al. 2022). Yet, as expectations for transparency and accountability rise, AI systems generally and automated systems more specifically also introduce new vulnerabilities when deployed within complex, multi-actor coordination settings (Behl et al. 2025).

In response, the notion of Meaningful Human Control (MHC) has become increasingly prominent, which according to Cavalcante Siebert et al. (2022) requires that: human–AI systems should (i) *track* relevant human reasoning and ethical considerations, and (ii) enable humans to *trace* system behaviour and consequences (Cavalcante Siebert et al. 2022). In practice, however, crises are characterized by severe time pressure, uncertainty, and fragmentation, leaving little room for deliberation or fact-checking under operating conditions (Comes 2024). Under these constraints, "tracing" is cognitively demanding: provenance inspection, interpreting model explanations and envisioning unexpected by-products compete with urgent coordination tasks and rapidly evolving operational risks. This creates a structural tension: the very situations that most demand human control are those in which cognitive bandwidth is scarcest and coordination links are most brittle (Wolbers et al. 2017; Schakel and Wolbers 2019). Without careful design, automation prospects may therefore shift authority without guaranteeing control, amplifying known vulnerabilities such as bias propagation and misalignment with local practices (Behl et al. 2025).

Amid these challenges, existing work has started to map *where* AI can add value but less so *how* to maintain system-wide control under time pressure (Shayganmehr et al. 2021). Particularly how to operationalize hybrid coordination at scale—when to defer, when to override, and how to preserve traceability—remains underdeveloped. Adjacent high-stakes domains reveal similar challenges: self-driving transport and process control emphasize human oversight but struggle with calibration of trust and mode awareness under automation surprises (Endsley 2022); and clinical decision support demonstrates benefits when recommendations are coupled with expert judgment, yet accountability and explanation still hinge on workflow design (Jowarder 2025).

Existing contributions offer qualitative guidance on social embedding and critical success factors (Shayganmehr et al. 2021), but the field still lacks a comparable basis for translating and quantifying how effects of allocations of automation authority, timing choices (e.g., foresight horizons), and trust conditions cascade through coordination networks to shape outcomes. To advance this, the concept proposed in this paper captures macro-level organizational dynamics central to crises while linking them to micro-level behaviours of Human–AI Interacting (HAI) agents. Making this concept operational, however, requires not only an appropriate structural representation but also a model of what it means for humans and AI systems to interact.

Modelling Human–AI Teaming

Human–AI Teaming (HAT) modelling must account for fundamental asymmetries in sensemaking and decision-making between human and AI teammates: humans draw on shared intentionality, tacit knowledge, and contextual interpretation, whereas AI systems optimise task-bounded objectives via statistical inference (Ilievski et al. 2025). These differences imply that AI should not be treated as a conventional teammate (Zhang et al. 2022). Levels of Automation (LoA) frameworks provide a useful taxonomy for degrees of which different kind of organisational functions can be automated (Endsley 2017). However, LoA has been criticized for behavioural unrealism and overlooking implementation challenges, limiting its prescriptive value in real teams (Jamieson and Skraaning 2017).

Since then, research on the behavioural dimensions of integrating AI into teams—across individual, team, and organisational levels—has expanded, with trust emerging as a central mechanism (Carter-Browne et al. 2021; O'Neill et al. 2020). Zhou, Duan, et al. (2025) demonstrate through controlled drone-coordination experiments how trust diffuses across different HAT configurations based on reliability, though this work has yet to be extended to larger or more variable systems. Moreover, Vaccaro et al. (2024)'s systematic review and meta-analysis of "when combinations of humans and AI are useful" shows that, on average, human–AI teams currently often perform significantly worse than the best human or AI alone. This is due to the heterogeneity in HAI performance driven by operational context, human background, and AI implementation. To generalise and quantify such behavioural insights across organisational levels, agent-based modelling (ABM) offers a natural next step, as it can capture emergent team states arising from micro-level interaction rules and heterogeneous roles (Will et al. 2020).

Crisis related ABMs already examine *human–human* information sharing and diffusion, but they typically do not represent AI as a distinct teammate with its own performance profile and trust dynamics. In humanitarian operations, Altay and Pal (2013) show how information hubs and information reliability shape diffusion and latency of response (Altay and Pal 2013). In hurricane risk communication, Watts et al. (2019) simulate how forecaster, media, public official, and peer networks shape protective actions; related work quantifies dissemination time distributions for official versus peer warnings using ABM (Siam et al. 2023). On the modelling process itself, Nespeca et al. (2023) provide a qualitative-to-ABM methodology and illustrate it for disaster information management, underscoring how ABM can be grounded in empirical inquiry (Nespeca et al. 2023). Together, these studies confirm that ABM can represent crisis information flows and decision diffusion under pressure, yet they stop short of modelling AI agents and human–AI trust calibration within organisational coordination networks—precisely where system-level effects and authority shifts emerge and should be explored.

Prior work has clearly clarified key elements of human–AI teaming and its relevance for crisis management. However, three critical gaps remain. First, there is a lack of system-level approaches that trace how automation choices cascade through coordination networks, calls to operationalize control principles in the complex, high urgency, high stakes settings that come with crisis management remain unanswered. Second, there is no shared abstraction foundation that enables comparable analyses across empirical studies in differing organizational contexts, thereby constraining cumulative cross-over insights. Third, there is little quantitative knowledge on how specific human–AI teaming dynamics affect performance at scale, as current crisis simulation models have not yet captured these intricacies. This paper directly addresses these gaps by proposing a conceptual model that supports high-level, structured representations of human–AI teaming configurations and enables scenario-based "what-if" simulations of alternative automation regimes and trust-mediated human–AI interaction settings in large-scale crisis management coordination. In doing so, it supports cross-context comparability and provides a means to explore operational guidance on when, how, and to what extent automation can enhance performance without undermining minimal meaningful human control.

CONCEPT DESIGN

Concept

Existing research offers valuable but fragmented insights into crisis information management, human–AI teaming, automation levels, and trust. Yet these strands rarely converge into a dynamic representation that captures the non-linear, networked, and trust-mediated interactions shaping coordination during crises. To address this gap, we propose a conceptual design that models Human–AI team interaction as a socio-technical system unfolding in time and space. We represent sensing, analysis, information sharing, and decision-making as distributed processes across heterogeneous decentralized teams, jointly influenced by automation and trust. Building on principles from organizational agent-based modelling and social-network approaches, the framework offers a minimal yet expressive structure that can be instantiated with domain-specific parameters or extended to suit diverse crisis contexts.

Overall Structure and Environment

Figure 1 depicts the model's three core components: the situational picture with its dynamic locations, the operational teams embedded within these locations, and the control-room teams maintaining oversight of the situational picture from outside of it. This setup builds on literature mapping crisis management coordination in strategic and operational teams within hierarchical settings (Comes et al. 2020; Abbas and Miller 2025) and was developed in consultation with AI COMPASS partners in crowd management. The situational picture consists of interconnected locations with time-dependent state variables, such as hazard intensity, sensor outputs, or system load. Each state variable can be linked to a configurable threshold used to mark elevated risk.

Operational teams operate at specific locations and obtain local observations while executing actions that directly modify the state of their current location. Their state observation is inherently partial and location-bounded, reflecting the limited situational access typical in field operations (Comes et al. 2020). Control-room teams, by contrast, have an aggregate view of all the locations and can therefore see the entire situational picture and if available take into account predicted states per location.

Information exchange between components shapes coordinated activity: operational teams report their local state indirectly to the control room, which integrates these inputs with its broader and, where applicable, forecast-enhanced perspective directly from the situational picture to generate action recommendations. Only operational teams can directly alter the environment through actions executed at their assigned locations. They may follow control-room recommendations or, under conditions in which local indicators exceed critical thresholds, choose to act autonomously based on their localized and partial observations. This mirrors decentralization patterns in crisis response (Schakel and Wolbers 2019).

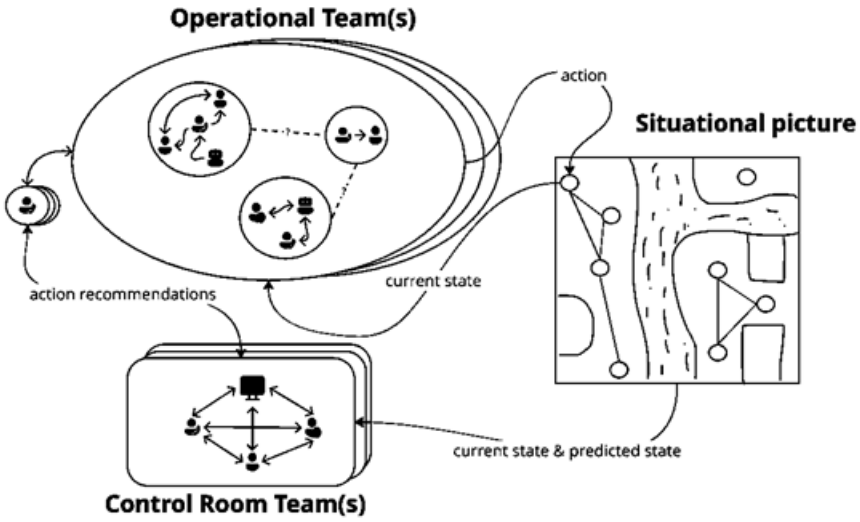


Figure 1. Concept Environment

Agent Types and Placement

Each agent at its core is conceptualized to possess one or more functional capabilities aligned with the information-management cycle: sensing, information analysis, sharing, and deciding (Lee et al. 2026). These can operate at three Levels of Automation—Human, Hybrid, Automated—following a simplified version of established LoA typologies (Endsley 2017; O'Neill et al. 2020). This combination is then set to a corresponding performance in terms of reliability and latency score (see Table 1). Here, latency connects directly to the extensive literature that treats timeliness and delays as key drivers of coordination (Van De Walle and Comes 2015; Nespeca et al. 2023), while reliability captures both direct performance (e.g., the accuracy of AI outputs) and broader definitions relevant to an agent's effectiveness in handling uncertain or possibly false information (Vaccaro et al. 2024). Reliability scores can thus reflect operational context, agent background, and explainability factors. These primary agent drivers are complemented by flow-directional *trust levels* for each connected agent per function.

Agents operate within a social-network layer linking operational teams and control-room staff across organizations (see example setting Figure 2). This social layer represents coordination protocols and hierarchical structures,

Table 1. Model setup reliability (r) and latency (l) concepts per function and level of automation

Function	$Human_r$	$Human_l$	$Hybrid_r$	$Hybrid_l$	$Automated_r$	$Automated_l$
Sensing (Se)	$r_{\text{Se,Hu}}$	$l_{\text{Se,Hu}}$	$r_{\text{Se,Hy}}$	$l_{\text{Se,Hy}}$	$r_{\text{Se,A}}$	$l_{\text{Se,A}}$
Information Analysis (IA)	$r_{\text{IA,Hu}}$	$l_{\text{IA,Hu}}$	$r_{\text{IA,Hy}}$	$l_{\text{IA,Hy}}$	$r_{\text{IA,A}}$	$l_{\text{IA,A}}$
Sharing (Sh)	$r_{\text{Sh,Hu}}$	$l_{\text{Sh,Hu}}$	$r_{\text{Sh,Hy}}$	$l_{\text{Sh,Hy}}$	$r_{\text{Sh,A}}$	$l_{\text{Sh,A}}$
Decision-Making (DM)	$r_{\text{DM,Hu}}$	$l_{\text{DM,Hu}}$	$r_{\text{DM,Hy}}$	$l_{\text{DM,Hy}}$	$r_{\text{DM,A}}$	$l_{\text{DM,A}}$

enabling the integration of agent-based modelling with social-network analysis to examine causal relationship between organisational structures and network dynamics (Will et al. 2020). Links mediate the exchange of information, after which agents may execute one or more of their available functions with some latency or refrain from acting. This decision is shaped jointly by the source of the information (e.g., the social relation to the sending agent in the form of a trustlevel) and its content (e.g., location state change), aligning with human–AI teaming principles identified by O'Neill et al. (2020). The underlying decision processes may range from simple threshold-based rules to more sophisticated models such as an Observe–Orient–Decide–Act (OODA) cycle (Abdollahian and Jeffries 2024).

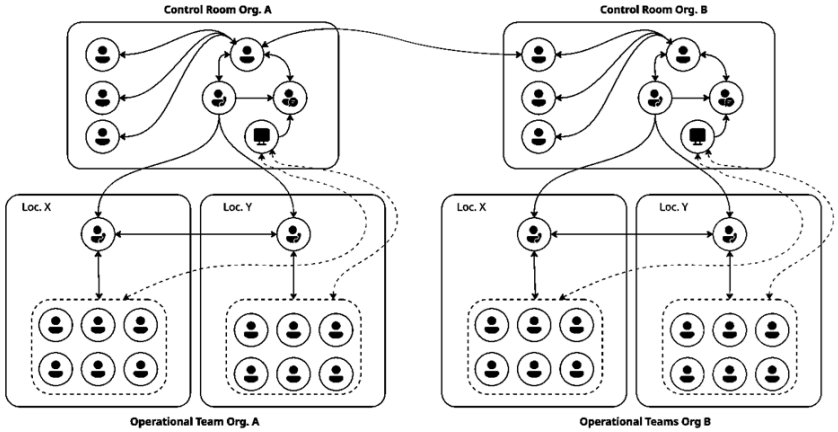


Figure 2. Concept Agent Network Structure

Automation, Performance, and Trust

The conceptual model then uses this basic setup—agents characterised by function-specific latency, reliability, and trust levels—to simulate how they interact with one another and with the evolving operational picture, as illustrated in Figure 3.

Three main dynamics are at play. First, when information arrives from another agent, the combination of Level of Automation (LoA) and the function of interest determines the baseline performance for the receiving agent in the requested function, as shown in Table 2. Latency and reliability capture how quickly and effectively the agent evaluates whether—and how—to perform its function, reflecting findings that information delays and information quality critically shape decision-making effectiveness and system responsiveness (e.g. (Lythreatis et al. 2025; O'Neill et al. 2020)). Once information reaches an operational agent with the function "Decision-Making", in this case acting, the timing influenced by the total latency and reliability of the path taken directly influence both the appropriateness of the resulting action and its outcome in the situational picture. This aligns with research on timing and sequencing in human–AI teaming reported by Zhou and Gorman (2024).

Second, the generated outcome produces a fitness score that reflects the agent's evaluation of its decision. This score can be computed through mechanisms that capture deviations between expected and actual outcomes, ranging from simple threshold-based satisfaction to more advanced approaches such as Bayesian updating (Chan and Adali 2012). These fitness assessments drive updates to trust levels toward the agents involved in producing or relaying the information. However, in cases of unexpected risk or sustained patterns of effective or ineffective control, the fitness signal can also be used to update the broader system of trust levels between agents, allowing trust to shift not only through direct interaction but also through inferred assessments of system-wide coordination (Zhou, Duan, et al. 2025; Ulfert et al. 2023).

Third, Trust updates feedback into the decision process influencing latency and reliability, as trust has been shown to mediate communication speed and willingness to act upon shared information (O'Neill et al. 2020). High trust generally accelerates information flow and increases compliance (Zhou, Duan, et al. 2025).

Finally, the conceptual model acknowledges that latency, reliability, and trust are sensitive to operational context and agent background (Vaccaro et al. 2024). For specific use cases, these parameters can be calibrated using empirical research on explainability, human–AI characteristics, and operational conditions. Alternatively, they can be further developed as variables of interest, allowing exploration of their effects—an approach well suited to agent-based modelling. For now, however, we treat them as optional parameters to preserve comparability and interpretability without introducing extra complexity

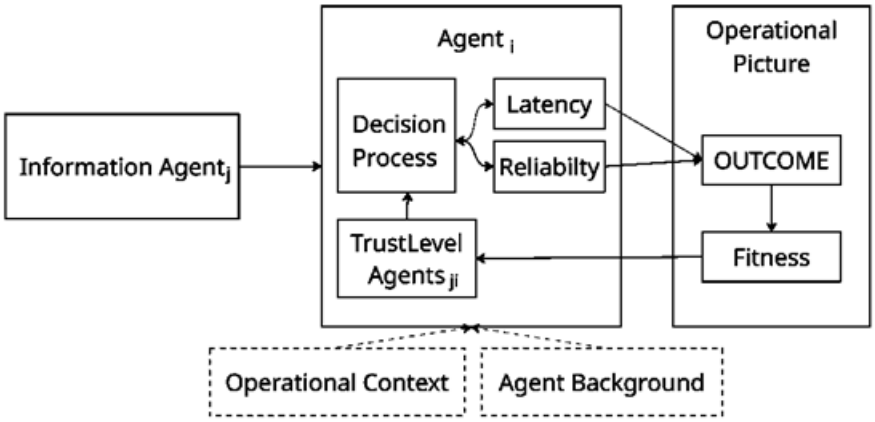


Figure 3. Concept Main Dynamics

Performance Outputs and Key Metrics

The concept integrates Social Network Analysis and Agent-Based Modelling to capture both agent-level performance and system-level emergent behaviour (Will et al. 2020). Process-oriented indicators, such as evolving trust patterns, align with the satisfaction and expectation dimensions in Rongier et al. (2012)'s framework, while ABM-derived situational outcomes capture Rongier et al. (2012)'s efficiency and relevance of agents' decisions with regards to their means, together shaping impact in terms of the effectiveness of the response coordination.

In context, satisfaction and expectation can for instance be operationalised through changes in trust levels over time and the extent to which events progress as planned, requiring minimal corrective actions. Efficiency could be represented by the degree to which objectives—such as maintaining control over (anticipated) crises—are achieved, while relevance pertains to whether additional means in coordination protocols such as AI-enabled decision support function appropriately without introducing unnecessary information clutter, incorrect recommendations, or additional workload.

Additionally the ABM can also log where decisions originate in the network and with which trust levels, what reliability and which speed they are made. This concept offers a means to start evaluating the *traceability* condition of (Cavalcante Siebert et al. 2022) Meaningful Human Control.

Minimal Model Application: Crowd Management

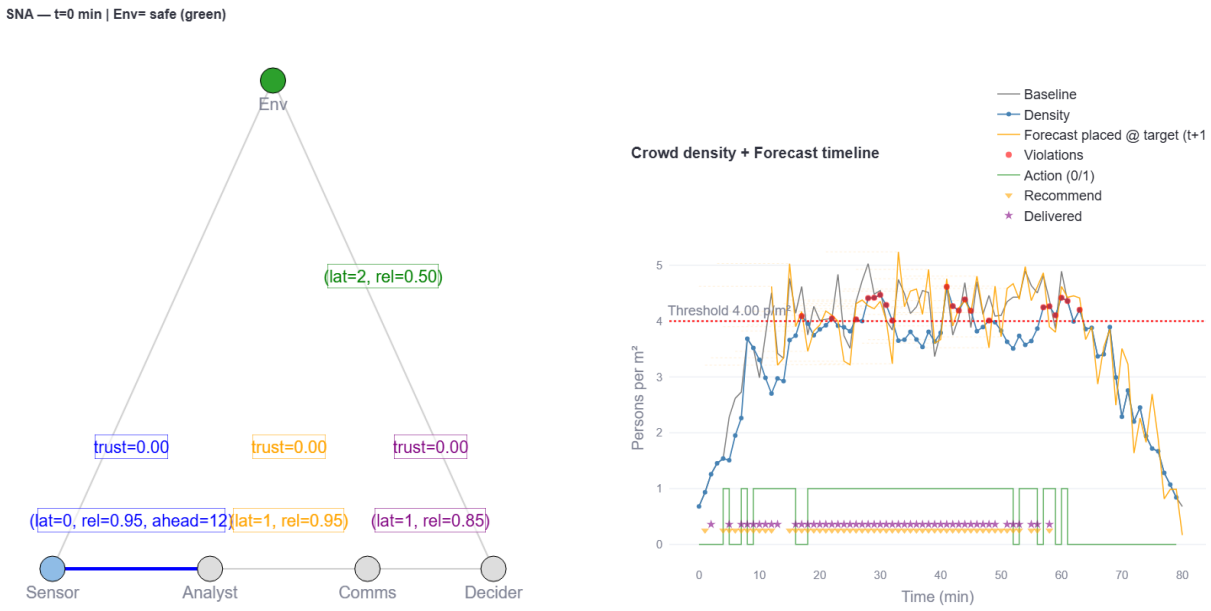
To illustrate the proposed concept, we implement a *minimal model* in a crowd-management case. Crowd management is crisis-adjacent: it involves continuous monitoring of evolving conditions, preferably pro-active but often reactive risk control, and distributed coordination between field and control-room roles—elements that mirror crisis response in other contexts (Feliciani et al. 2021). Accordingly, the model's operational objective is straightforward: track local crowd density and keep it below a predefined risk threshold through timely recommendations and actions.

The setup is intentionally simple. A pipeline of four agents aligned with the information-management cycle—*Sensing*, *Analysis*, *Sharing*, and *Decision*—interacts over a social-network link structure . This minimal structure is sufficient to expose how *levels of automation*, *trust*, *latency*, and *reliability* jointly determine whether density is controlled early (anticipation) or only after escalation (reaction), without claiming venue-specific predictions.

Environment and Agents

The minimal instance models a pipeline-style management network linked to a single observable location (e.g., a concourse), where crowd density x_t is simulated in discrete one-minute steps. The ground-truth (no-intervention) density is instantiated as a stylised event timeline with *ingress* (arrival and fill), *circulation* (steady occupancy with local movements), and *egress* (clear-out), a pattern widely documented in pedestrian and crowd operations (Still et al. 2020). Surges approximate short-lived arrival spikes (e.g., a train unload), and state-dependent noise reflects the rise in variability and instability at higher densities. A global risk threshold is set at 4 persons/m^2, consistent with level-of-service reasoning that links density to flow and safety (Feliciani et al. 2021).

The pipeline is instantiated with four agents, each aligned to one function of the information-management cycle, where only the Sensor and Decider interface directly with the environment (as shown in Figure 4a). Each agent has a designated Level of Automation (Human, Hybrid, Automated), initial trust on adjacent links, and performance parameters inherited from the LoA per function in Table 2. These default values are informed by interactions with crowd management researchers and practice partners of the project. They reflect a broad consensus that sensing (detection/forecasting) and information sharing benefit from higher automation, whereas analyzing and deciding/acting often gain from stronger human involvement (Lauras and Comes 2015; Van De Walle and Comes 2015; Lythreatis et al. 2025; Behl et al. 2025; Kordi and Ertz 2025). The Sensor provides a noisy estimate $\widehat{x}_t$ or a Δt-ahead forecast. Noise increases with lower reliability, larger forecast horizons and cases of ground truth peak behaviour. This consistent with AI detection and forecasting challenges (Kordi and Ertz 2025).



(a) Social Network Structure Minimal Model **(b) Situational Picture Minimal Model**

Figure 4. Minimal Model Visualisation

Table 2. Default reliability (r) and latency (l) values per function and level of automation. Latency is measured in minutes; Reliability represents the percentage of correct output

Function	$Human_r$	$Human_l$	$Hybrid_r$	$Hybrid_l$	$Automated_r$	$Automated_l$
Sensor	0.40	5	0.70	3	0.95	0
Analyst	0.80	3	0.90	2	0.70	0
Comms	0.60	2	0.75	1	0.90	0
Decider	0.90	4	0.85	3	0.70	0

Trust Dynamics

Trust is updated after each timestep using a simple outcome–bandwidth rule. A step is *timely-correct* when a recommendation-based action arrives in time and the density remains within a controlled bandwidth around the risk threshold, or when no recommendation/action is issued and the density remains controlled (a correct non-alarm). It

is *late-or-incorrect* when an unnecessary action is taken (the current density does not remain within the bandwidth), when a recommendation arrives too late or not at all and a breach outside the bandwidth follows, or when a local override/reflex action is triggered at the decider because no timely, trusted recommendation was available. Directed link trust between agents at time t, denoted $\tau_{ij}(t)$, then evolves as:

$$\tau_{ij}(t+1) = \text{clip}_{[-1,1]}\big(\tau_{ij}(t) + 0.1 * \mathbf{1}\{\text{timely-correct}\} - 0.3 * \mathbf{1}\{\text{late-or-incorrect}\}\big).$$

Here the asymmetric learning rates capture the widely observed *faster loss than gain* of trust, and the bounding reflects that trust saturates at plausible limits (Zhou, Duan, et al. 2025; Duan et al. 2025). Trust then feeds back into effective reliability (r_j^{eff}) and latency (ℓ_j^{eff}) for the receiving agent j when processing information arriving along link ($i \to j$), given j's earlier reliability (r_j) and latency (ℓ_j):

$$r_j^{\text{eff}} = \min\{1,\ r_j + 0.5\,\tau_{ij}\}, \qquad \ell_j^{\text{eff}} = \max\{0,\ \ell_j - \tau_{ij}\},$$

This captures how positive collaboration histories reduce coordination friction, while eroded trust slows coordination and increases the likelihood of ignoring advice (O'Neill et al. 2020)

Key Processes and Outputs

By instantiating the concept in a compact pipeline, the model enables controlled exploration of how automation and trust shape coordination performance. Situational outcomes quantify safety in terms of risk threshold breaches and operational cost in terms of the number of actions (a_t) taken over the duration of the event (T):

$$KPI_{\text{breach}} = \sum_{t=1}^{T} \mathbb{I}(x_t \geq 4) \qquad KPI_{\text{action}} = \sum_{t=1}^{T} a_t.$$

Process outcomes track coordination quality by change of trust levels and risk of emerging fragmentation by the number of ignored recommendations (R_t^{ignored}):

$$KPI_{\Delta\tau} = \sum_{(i,j)\in E} \left(\tau_{ij}^{T} - \tau_{ij}^{0}\right) \qquad KPI_{\text{ign}} = \sum_{t=1}^{T} R_t^{\text{ignored}}.$$

Model Flow

Each run is initialized with the scenario specification (LoA by function), a forecast horizon, link-specific initial trust, the environment process (density, inflow, noise), and KPI counters. At each timestep $t \in \{1, \ldots, T\}$ the simulation proceeds as follows.
(1) *Deliver queued messages*: all messages whose delivery time equals t are released.
(2) *Sense*: the Sensor produces an observation or $+\Delta t$ steps ahead forecast of the density and sends it to the Analyst with reliability and latency determined by its LoA setting.
(3) *Analyse*: upon receipt, the Analyst compares the signal to an alert threshold (3.5 persons/m^2). If exceeded, it issues a recommendation and forwards it to Comms with Sensor-Analyst trust modulated effective latency/reliability.
(4) *Share*: Comms relays any recommendation to the Decider, again under trust- and LoA-dependent latency/reliability, which in this case are messages being delayed or lost.
(5) *Decide/Act*: the Decider acts with his LoA instructed latency if either (i) a sufficiently trusted recommendation arrives based on his trust modulated reliability or (ii) a local override threshold is seriously breached (4.2 persons/m^2), capturing reactive fragmentation. An action reduces current density by 10% and suppresses inflow by 90% for a short duration. Ignored recommendations are recorded when advice is present but no action is taken.
(6) *Update state and KPIs*: the environment is advanced using inflow, noise, and any action effects; breaches ($x_t \geq 4$ persons/m^2) and actions are tallied.
(7) *Update trust*: system outcomes are evaluated as "timely-correct" or "late-or-incorrect" and link trust is adjusted accordingly. Trust then feeds back into the next step's effective latency and reliability.
After T steps, KPI aggregates (breaches, actions, ignored recommendations, and total trust change) are returned.

Experimental Design

The experimental design explores the minimal model's behaviour as a proof of concept for the conceptual model design. Rather than seeking optimisation, it uses illustrative scenarios to generate qualitatively interpretable patterns that can be compared to the literature. Five configurations (Table 3) vary levels of automation regimes, forecast horizons, and initial trust, enabling comparison of contrasting coordination regimes and their effects on coordination flow, reliance, and emergent team behaviour.

Table 3. Scenario configurations used in the experimental design

Scenario	*LoA Regime*	*Forecast Horizon*	*Initial Trust*
Baseline Coordination	Hu–Hu–Hu–Hu	0 min	+0.5
Automation (Reactive)	A–Hy–Hy–Hu	0 min	+0.5
Hybrid Foresight (High trust)	A–Hy–Hy–Hu	12 min	+0.5
Hybrid Foresight (Low trust)	A–Hy–Hy–Hu	12 min	–0.5
Automation-driven Foresight	A–A–A–Hy	12 min	+0.5

RESULTS

Figure 5 summarises the KPI distributions across scenarios given 1000 Monte-Carlo per scenario. The results show that *Reactive Automation* (orange in Figure 5, current operational norm) performs most robustly overall: it yields few risk breaches, maintains a moderate action load, minimises unnecessary recommendations, and produces a positive trust trajectory. *Hybrid Foresight (High Trust)* achieves comparable breach and workload outcomes but generates a large number of unnecessary recommendations, likely contributing to its declining trust—consistent with findings that high initial trust can promote premature compliance (Zhou, Duan, et al. 2025). The *Base Case* performs relatively poorly, aligning with evidence that reactive coordination structures benefit markedly from data-driven decision support (Lauras and Comes 2015). *Full Automation* does not outperform human-involved variants under the present reliability and latency assumptions, exhibiting more risk breaches than Hybrid or Reactive Automation. Interestingly, *Hybrid Foresight (Low Trust)* is—alongside Reactive Automation—the only setting that produces increasing trust over time. This suggests that lower initial trust may, in some contexts, initiate a virtuous rather than vicious trust–performance feedback loop, echoing recent work on human–AI calibration dynamics (Duan et al. 2025).

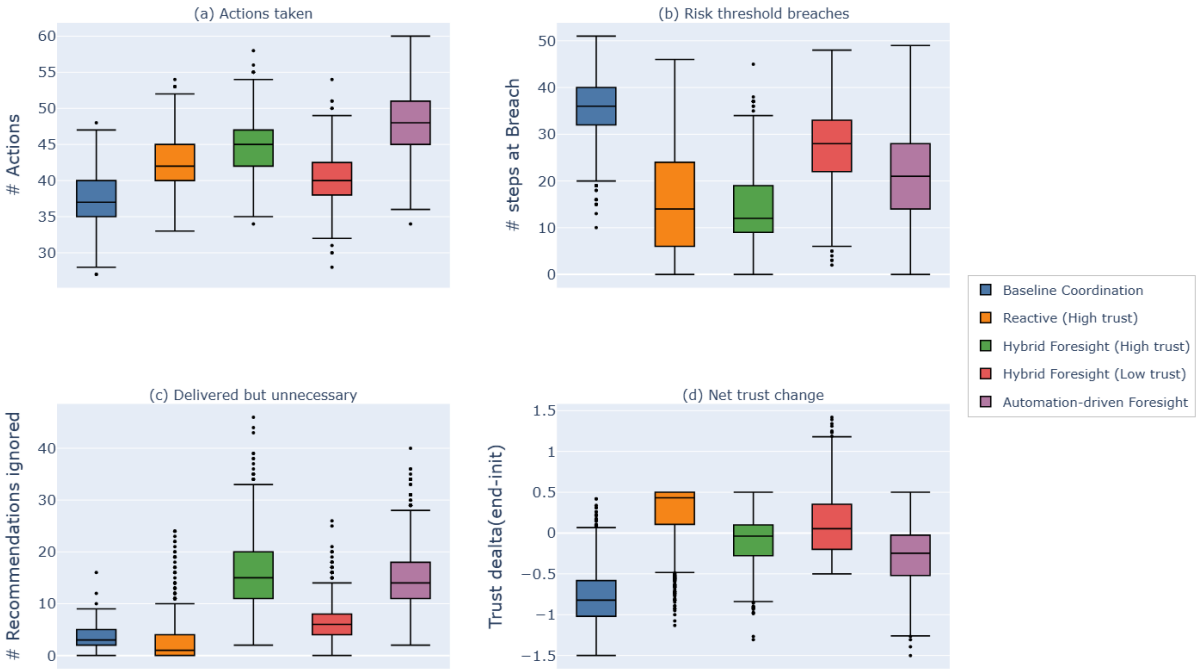


Figure 5. KPI's Generic Statistics

When taking a wider view of the interquartile ranges, 90% confidence whiskers, and outliers, several additional points emerge. First, the confidence intervals for *Risk Breaches* and *Actions Taken* are broad across all scenarios,

indicating strong sensitivity to the model's stochastic components—which is mostly governed by the reliability scores. This suggests that when reliability falls favourably, outcomes are highly effective; when it does not, performance can deteriorate sharply. This aligns with findings on timing and sequencing effects (Zhou and Gorman 2024) and with crisis-management theories of non-linear escalation (Helbing and Mukerji 2012).Notably, the Reactive Automation regime, despite strong mean performance, shows a wide interquartile range in risk-threshold breaches, highlighting the need for large-scale simulation under diverse conditions to understand when the system shifts between performance modes. Finally, the high upper whisker of net trust change in the Hybrid Foresight (Low Trust) scenario indicates that substantial trust growth can occur despite an update rule that penalises failures more strongly than successes—suggesting a dynamic that warrants further study.

Given the wide confidence intervals, a significance test was added to assess whether the scenarios perform significantly differently from each other. The KPIs were non-normal with unequal variances; we therefore used two-sided Mann–Whitney U tests. All pairwise differences were significant ($p < 0.001$) except for Hybrid–Foresight (Low Trust) vs. Automation–Driven Foresight on the 'delivered-but-no-reason' KPI, indicating generally robust scenario effects.

Situational performance

Figure 6 shows the average realised density per scenario, with the "no intervention" line in black. Foresight and levels of automation—mediated by trust—jointly determine how early the system responds to rising density. Baseline coordination tracks the ground truth and exceeds the threshold for prolonged periods, as expected. Hybrid Foresight (High Trust) provides the strongest anticipatory control, keeping density well below the threshold. With Low Trust, forecasts arrive but translate into delayed action, producing a higher mid-event peak. Although Low Trust yields smaller breaches (lower immediate risk), it reflects fragmented, late-stage interventions that work quickly yet occur after escalation—consistent with Wolbers et al. (2017).

Beyond the summary statistics, the timeline shows that Hybrid Foresight (High Trust)—despite many currently deemed unnecessary actions from 5— has much better stabilisation (at the risk recommendation level) in comparison to all other scenarios, a potentially interesting dynamic. Also notably, all foresight-based scenarios, especially Hybrid Foresight (High Trust), show a distinctive threshold breach at the end of the plateau (around t=60), this could indicate a bad forecast (anticipating an earlier decline) and raises Cavalcante Siebert et al. (2022)'s question of meaningful human control: when did a human last shape the decision? In the automation scenario, possibly never, as only the decider remains hybrid with limited situational awareness under high risk. In hybrid foresight scenarios, the person who should have control can be traced to analyst, whose effectiveness depends on latency, reliability, and current trust. Reactive Automation lacks this late bump, suggesting control trade-off dynamic introduced by foresight.

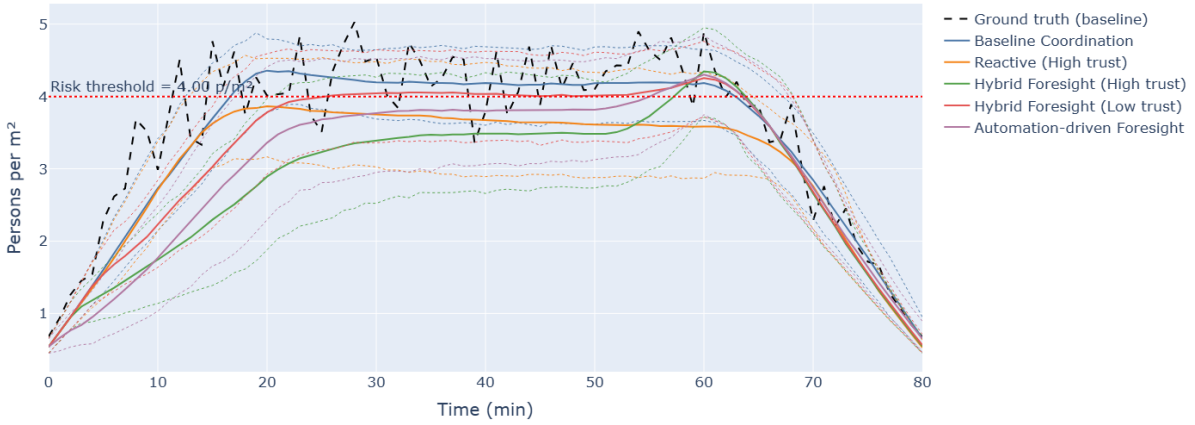


Figure 6. Simulated Performance: Density Time Series

Process performance Analyses

Figure 7 shows how the operational decider agent's trust level evolves over time. The dynamics align with earlier findings: Baseline Coordination steadily loses trust as late, inaccurate cues accumulate. Automation (Reactive) shows mild but consistent trust gains, driven by accurate measurement and a relatively fast pipeline. High-trust

foresight exhibits an early dip—likely due to over-recommendations that deviate below the ground truth—causing overstimulation and trust erosion by bad timing (Zhou and Gorman 2024). Low-trust foresight, by contrast, trends upward modestly as useful recommendations are confirmed over time. Notably, before the late plateau miss, all foresight trajectories show partial stabilisation, suggesting a local equilibrium. This matters analytically: once trust stabilises, over-reliance and AI-use bias become easier to monitor and audit, connecting to the second condition of meaningful human control (Cavalcante Siebert et al. 2022).

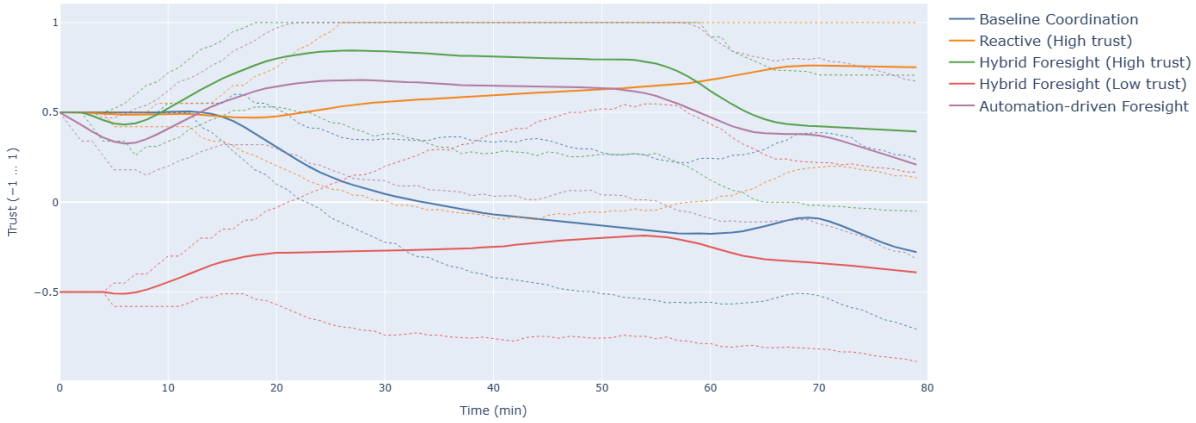


Figure 7. Simulated Trust Evolution (decider agent)

DISCUSSION

This work represents an initial step toward developing a systematic representation of Human–AI interaction dynamics in crisis management (CM). By proposing a concept agent-based model (ABM) that captures core mechanisms of trust formation, information exchange, and coordination between human and artificial agents, we aim to offer a conceptual scaffold that can support more comprehensive future modelling efforts.

The minimal model produced results that show interesting dynamics which are also broadly in line with patterns reported in existing literature on coordination dynamics and distributed decision-making in emergency contexts. While this validation is limited in scope, the combination of conceptual alignment with literature and early empirical resonance with stakeholders suggests that the underlying logic of the concept is promising and sufficiently robust to motivate the development of larger-scale exploratory simulations.

Strengths and Limitations

A key implementation strength of the framework is its high transferability. Because it is structured around generalizable mechanisms—including trust updating, information accuracy, role responsibilities, and communication networks—it can, in principle, be applied across sectors, operational goals, and organizational configurations. This characteristic aligns with the adaptability observed in similar ABM approaches, such as those used in crisis management but was before not extended to Human-AI interaction.

A central implementation limitation is that the precise functional forms and parameters governing trust updating, human–AI reliability perception, and escalation behavior are not yet fully empirically established. As such, the current model remains conceptual rather than prescriptive. Nonetheless, it is reasonable to expect that these parameters can be refined through serious gaming environments, controlled experiments, empirical observations during exercises, or analyses of crisis case studies(Ahrweiler et al. 2024; Kordi and Ertz 2025; Zhou, Duan, et al. 2025).

A second limitation concerns conceptual abstraction. The current model intentionally omits individual traits or role-specific identities to maintain comparability and analytical tractability. However, crises often escalate due to individual human behaviour, cognitive overload, or deviations from protocol—factors linked to personal characteristics, experience levels, or situational stress rather than purely structural elements. Incorporating such heterogeneity would improve realism. Future expansions of the concept could incorporate dynamic conceptual implementation agent background factors and operational contextual. Note, however, that maintaining comparability and analytical tractability requires at least a shared agreement on the underlying dynamics, which is challenging to establish across such a broad and heterogeneous body of (empirical) research (Vaccaro et al. 2024).

From a technical perspective, the current minimal model is functional but primarily illustrative. Two major development trajectories address this limitation:

1. **An explorative, mid-scale model:** This version would incorporate multiple locations, heterogeneous agents, and diverse decision functions. It could support investigations into core CM questions such as what tasks should be automated, when, and to what extent. From here it could provide insight into resource allocation strategies and help design protocols that ensure meaningful human control.
2. **A high-resolution model:** Enabled by increasing data availability, future work may leverage communication logs, historic crisis data, and digital-twin infrastructures to construct fine-grained interaction networks. Such models may eventually learn realistic cause–effect relations and support predictive or prescriptive analytics for CM operations.

Future Research

Advancing this work toward real-world application will require a combination of methodological development and empirical investigation. Several promising directions emerge from the current concept when implemented at scale. First, scenario discovery offers a systematic way to explore system behaviour under diverse and uncertain future conditions. This is particularly relevant for examining deep uncertainties surrounding trust evolution, human–AI interaction patterns, and variations across different crisis case settings(Kwakkel 2015). Second, intervention strategy testing can support structured evaluation of training approaches, automation choices, and resource allocation policies. For example, organisations might experiment with trust-engineering training to enhance human readiness for AI-supported operations (Ezer et al. 2019), or explore the use of automated signalling systems during large events or festivals to improve coordination. Third, system-level social network analyses can help uncover leverage points, unintended consequences, and emergent phenomena within human–AI teaming dynamics. Such insights can be directly linked to the operationalisation of Cavalcante Siebert et al. (2022)'s meaningful human control and ultimately inform the development of future-proof crisis management protocols.

Finally, as this paper presents work in progress, we encourage readers to engage directly with the minimal model and explore its behaviour. The full formulation and interactive dashboard of the minimal model—with all parameters dynamically adjustable—are openly available through the public Github Repository (Dert 2026). We warmly welcome comments, suggestions, and creative ideas for extending the concept and its dynamics.

CONCLUSION

This paper introduces a conceptual model and a minimal agent-based implementation to explore how automation choices, trust dynamics, and coordination structures interact in crisis information management. The results suggest that, even with a highly abstracted setup, the model produces several widely observed and theoretically grounded Human-AI teaming patterns. These findings indicate the potential of this concept to support larger-scale, systematic "what-if" analyses in cross-crisis-context-comparable settings, contributing toward the objective of establishing a transferable foundation for studying human–AI teaming at the system level for crisis management.

ACKNOWLEDGEMENTS

This work was funded by the Dutch Research Council (NWO) under KICH1.VE04.22.007 and is part of the AI-COMPASS project. Furthermore we would like to thank all contributing consortium partners.

REFERENCES

Abbas, R. and Miller, T. (Mar. 2025). "Exploring communication inefficiencies in disaster response: Perspectives of emergency managers and health professionals". In: *International Journal of Disaster Risk Reduction* 120, p. 105393.

Abdollahian, M. and Jeffries, C. (2024). "Simulating Boyd's OODA Loop: towards an ABM of human agency and sensemaking in dynamic, competitive environments". In: ACHI 2024 : The Seventeenth International Conference on Advances in Computer-Human Interactions.

Ahrweiler, P., Gilbert, N., Juranyi, Z., Bicket, M., Coll, A. S., Kampis, G., Capellas, B. L., and Wurster, D. (May 2024). "Using ABM and Serious Games to Create "Better AI"". In: Annual Modeling and Simulation Conference (ANNSIM), pp. 1–16.

Altay, N. and Pal, R. (Aug. 2013). "Information Diffusion among Agents: Implications for Humanitarian Operations". In: *Production and Operations Management* 23.6, pp. 1015–1027.

Behl, A., Bhardwaj, S., Jayawardena, N., Pereira, V., and Roohanifar, M. (Dec. 2025). "Grass is always dark(er) on the other side: Exploring the dark side of artificial intelligence humanitarian supply chain operations". In: *Technological Forecasting and Social Change* 224, p. 124484.

Carter-Browne, B., Paletz, S., Campbell, S., Carraway, M., Vahlkamp, S., Schwartz, J., and O'Rourke, P. (June 2021). *There is No "AI" in Teams: A Multidisciplinary Framework for AIs to Work in Human Teams*. Tech. rep.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., et al. (May 2022). "Meaningful human control: actionable properties for AI system development". In: *AI and Ethics* 3.1, pp. 241–255.

Chan, K. and Adali, S. (Mar. 2012). "An agent based model for trust and information sharing in networked systems". In: IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, pp. 88–95.

Comes, T. (Feb. 2024). "AI for crisis decisions". In: *Ethics and Information Technology* 26.1.

Comes, T., Van De Walle, B., and Van Wassenhove, L. (June 2020). "The Coordination-Information Bubble in Humanitarian Response: theoretical foundations and empirical investigations". In: *Production and Operations Management* 29.11, pp. 2484–2507.

Dert, T. (Mar. 2026). *Crisis Management CP1*. https://github.com/Tamara653/Crisis_Management_CP1.

Duan, W., Flathmann, C., McNeese, N., Scalia, M. J., Zhang, R., Gorman, J., Freeman, G., Zhou, S., Hauptman, A. I., and Yin, X. (Apr. 2025). "Trusting Autonomous Teammates in Human-AI Teams - A Literature Review". In: CHI conference, pp. 1–23.

Dubey, R., Bryde, D. J., Dwivedi, Y. K., Graham, G., and Foropon, C. (Aug. 2022). "Impact of artificial intelligence-driven big data analytics culture on agility and resilience in humanitarian supply chain: A practice-based view". In: *International Journal of Production Economics* 250, p. 108618.

Endsley, M. R. (Oct. 2017). "Level of automation forms a key aspect of autonomy design". In: *Journal of Cognitive Engineering and Decision Making* 12.1, pp. 29–34.

Endsley, M. R. (Nov. 2022). "Supporting Human-AI Teams:Transparency, explainability, and situation awareness". In: *Computers in Human Behavior* 140, p. 107574.

Ezer, N., Bruni, S., Cai, Y., Hepenstal, S. J., Miller, C. A., and Schmorrow, D. D. (Nov. 2019). "Trust Engineering for Human-AI teams". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63.1, pp. 322–326.

Feliciani, C., Shimura, K., and Nishinari, K. (Jan. 2021). *Introduction to crowd management*.

Helbing, D. and Mukerji, P. (Jan. 2012). "Crowd Disasters as Systemic Failures: Analysis of the Love Parade Disaster". In: *SSRN Electronic Journal*.

Hempel, L., Kraff, B. D., and Pelzer, R. (Apr. 2018). "Dynamic interdependencies: Problematising criticality assessment in the light of cascading effects". In: *International Journal of Disaster Risk Reduction* 30, pp. 257–268.

Ilievski, F., Hammer, B., Van Harmelen, F., Paassen, B., Saralajew, S., Schmid, U., Biehl, M., Bolognesi, M., Dong, X. L., Gashteovski, K., et al. (Sept. 2025). "Aligning generalization between humans and machines". In: *Nature Machine Intelligence* 7.9, pp. 1378–1389.

Jamieson, G. A. and Skraaning, G. (Oct. 2017). "Levels of Automation in Human Factors Models for Automation Design: Why we might consider throwing the baby out with the bathwater". In: *Journal of Cognitive Engineering and Decision Making* 12.1, pp. 42–49.

Jowarder, R. A. (Mar. 2025). "The Ethics of AI Decision-Making: balancing automation, explainable AI, and human oversight". In: *International Journal of Science and Research Archive* 14.3, pp. 435–443.

Kordi, M. and Ertz, M. (Sept. 2025). "Deciphering technological advancements for efficient disaster management and community resilience". In: *Technology in Society* 84, p. 103057.

Kwakkel, J. (2015). *EMA Workbench documentation — Exploratory Modeling Workbench*.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (Feb. 2021). "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research". In: *Artificial Intelligence* 296, p. 103473.

Lauras, M. and Comes, T. (Sept. 2015). "Special issue on Innovative Artificial Intelligence Solutions for Crisis Management". In: *Engineering Applications of Artificial Intelligence* 46, pp. 287–288.

Lee, C.-C., Comes, T., Finn, M., Pak, H., Hsu, C.-W., and Mostafavi, A. (Jan. 2026). "Roadmap toward Responsible AI in Crisis Resilience and Management". In: *IEEE Access* 14, pp. 11200–11215.

Lythreatis, S., Acikgoz, F., and Yassine, N. (Nov. 2025). "Artificial intelligence in humanitarian aid: A review and future research agenda". In: *Technovation* 151, p. 103415.

Nespeca, V., Comes, T., and Brazier, F. (Jan. 2023). "A methodology to develop Agent-Based models for policy support via qualitative inquiry". In: *Journal of Artificial Societies and Social Simulation* 26.1.

O'Neill, T., McNeese, N., Barron, A., and Schelble, B. (Oct. 2020). "Human–Autonomy Teaming: A review and analysis of the Empirical literature". In: *Human Factors The Journal of the Human Factors and Ergonomics Society* 64.5, pp. 904–938.

Rongier, C., Lauras, M., Galasso, F., and Gourc, D. (June 2012). "Towards a crisis performance-measurement system". In: *International Journal of Computer Integrated Manufacturing* 26.11, pp. 1087–1102.

Schakel, J. K. and Wolbers, J. (Dec. 2019). "To the edge and beyond: How fast-response organizations adapt in rapidly changing crisis situations". In: *Human Relations* 74.3, pp. 405–436.

Shayganmehr, M., Gupta, S., Laguir, I., Stekelorum, R., and Kumar, A. (Nov. 2021). "Assessing the role of industry 4.0 for enhancing swift trust and coordination in humanitarian supply chain". In: *Annals of Operations Research* 335.3, pp. 1053–1085.

Siam, M., Lindell, M. K., and Wang, H. (Dec. 2023). "Modeling of multi-hazard warning dissemination time distributions: An agent-based approach". In: *International Journal of Disaster Risk Reduction* 100, p. 104207.

Still, K., Papalexi, M., Fan, Y., and Bamford, D. (Apr. 2020). "Place crowd safety, crowd science? Case studies and application". In: *Journal of Place Management and Development* 13.4, pp. 385–407.

Ulfert, A.-S., Georganta, E., Jorge, C. C., Mehrotra, S., and Tielman, M. (Apr. 2023). "Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework". In: *European Journal of Work and Organizational Psychology* 33.2, pp. 158–171.

Vaccaro, M., Almaatouq, A., and Malone, T. (Oct. 2024). "When combinations of humans and AI are useful: A systematic review and meta-analysis". In: *Nature Human Behaviour* 8.12, pp. 2293–2303.

Van De Walle, B. and Comes, T. (Jan. 2015). "On the Nature of Information Management in Complex and Natural Disasters". In: *Procedia Engineering* 107, pp. 403–411.

Watts, J., Morss, R. E., Barton, C. M., and Demuth, J. L. (Sept. 2019). "Conceptualizing and implementing an agent-based model of information flow and decision making during hurricane threats". In: *Environmental Modelling Software* 122, p. 104524.

Will, M., Groeneveld, J., Frank, K., and Müller, B. (Feb. 2020). "Combining social network analysis and agent-based modelling to explore dynamics of human interaction: A review". In: *Socio-Environmental Systems Modeling* 2, p. 16325.

Wolbers, J., Boersma, K., and Groenewegen, P. (Aug. 2017). "Introducing a fragmentation perspective on coordination in crisis management". In: *Organization Studies* 39.11, pp. 1521–1546.

Zhang, Q., Lee, M. L., and Carter, S. (Apr. 2022). "You complete me: Human-AI teams and complementary expertise". In: *CHI Conference on Human Factors in Computing Systems*, pp. 1–28.

Zhou, S., Duan, W., Yin, X., Scalia, M., Hao, R., Weng, N., Funke, G., Tolston, M., Freeman, G., Schelble, B., et al. (Oct. 2025). "The spread of trust and distrust in human-AI teams". In: *Applied Ergonomics* 130, p. 104648.

Zhou, S. and Gorman, J. C. (Aug. 2024). "The impact of communication timing and sequencing on team performance: A Comparative study of Human-AI and All-Human teams". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 68.1, pp. 1769–1774.