Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Perception Uncertainty and Collective Robustness in AI-Driven Swarm UAV Systems

Oleksandr Kravchuk

oleksandr.kravchuk@pimee.ua

G.E. Pukhov Institute for Modelling in Energy Engineering, National Academy of Sciences

Artem Taranowski

G.E. Pukhov Institute for Modelling in Energy Engineering, National Academy of Sciences

Dmytro Sinko

G.E. Pukhov Institute for Modelling in Energy Engineering, National Academy of Sciences

Hryhoriy Kravtsov

G.E. Pukhov Institute for Modelling in Energy Engineering, National Academy of Sciences

Victor Samoylov

G.E. Pukhov Institute for Modelling in Energy Engineering, National Academy of Sciences

Research Article

Keywords: swarm intelligence, multiagent consensus, distributed stability, geospatial perception, AI risk modeling, cascading error amplification, autonomous UAVs, spectral graph theory, Analytic Hierarchy Process

Posted Date: March 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9065913/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Perception Uncertainty and Collective Robustness in AI-Driven Swarm UAV Systems

Oleksandr Kravchuk [1*], Artem Taranowski [2], Dmytro Sinko [3], Hryhoriy Kravtsov [4] and Victor Samoylov [5]

[1] oleksandr.kravchuk@pimee.ua
[2] artem.taranowski@pimee.ua
[3] dmytro.sinko@pimee.ua
[4] hryhoriy.kravtsov@pimee.ua
[5] viktor.samoilov@pimee.ua
[1,2,3,4,5] G.E. Pukhov Institute for Modelling in Energy Engineering, National Academy of Sciences of Ukraine
* Corresponding author: Oleksandr Kravchuk, oleksandr.kravchuk@pimee.ua;

Abstract

This paper examines risk formation and propagation in AI-driven swarm-based unmanned aerial vehicle (UAV) systems operating in complex environments. In distributed multiagent systems, perception errors made by individual agents can propagate through collective interaction and amplify at the swarm level. The lack of a structured approach to assessing such risks increases uncertainty during the preparation, training, and operational deployment of AI-based Geospatial Perception Models (AGPM).

A methodological framework is introduced to classify risks across the AGPM lifecycle, including knowledge selection, knowledge analysis, model training, and operational stages. The framework considers both human and technological factors and applies the Analytic Hierarchy Process (AHP) to assess their relative significance within a hierarchical multiagent setting. Particular attention is paid to cascading error propagation and its impact on collective swarm behavior.

Experiments with contemporary general-purpose multimodal AI systems were conducted to evaluate their ability to interpret graphical representations of complex geospatial scenes relevant to autonomous navigation. The results suggest that while semantic recognition remains stable, spatial-geometric reasoning and formal uncertainty estimation are limited. In swarm-based systems, these limitations may contribute to the amplification of local perception errors.

Overall, the analysis indicates that technological risks during operation, especially cascading error propagation among agents, have the strongest influence on the overall risk profile. The framework is intended to clarify how vulnerabilities emerging at different lifecycle stages may affect collective behavior in AI-driven swarm systems.

Keywords: swarm intelligence; multiagent consensus; distributed stability; geospatial perception; AI risk modeling; cascading error amplification; autonomous UAVs; spectral graph theory; Analytic Hierarchy Process;

1. Introduction

Swarm systems of UAVs represent a class of complex multi-agent systems [1]. According to their structural organization and control, they can be classified into four main types: centralized, hierarchical, distributed, and emergent swarming. Each type is characterized by specific principles of agent coordination, distribution of functions and responsibilities, as well as different levels of autonomy of individual agents within the swarm [2]. A distinctive feature of swarm systems is the emergence of collective behavior from local interactions between agents. In such distributed architectures, local perception errors or decision inconsistencies

may not remain isolated. Instead, they can propagate through coordination mechanisms and amplify at the collective level, potentially leading to large-scale behavioral deviations. This makes it important to examine the sources of such vulnerabilities in autonomous swarm systems.

Artificial intelligence (AI) plays an increasingly important role in modern UAV systems, particularly in swarm configurations [3]. However, the integration of AI into distributed autonomous systems remains methodologically and technically challenging. Swarm systems operate with high autonomy and nonlinear interaction dynamics, and their performance strongly depends on the reliability of the underlying AI algorithms.

While AI enhances scalability, adaptability, and mission effectiveness, it also introduces new risks [4]. These risks arise from technical factors (such as model architecture, training data quality, and computational resources), as well as from human factors (including dataset preparation, architectural decisions, and supervision of model behavior) [4]. This study focuses on risks associated with the preparation and training of AI models for UAV swarm systems. Hardware-related risks and platform-specific technological issues are outside the scope of this work.

2. Problem Statement

There is currently no dedicated and formalized methodology for assessing risks associated with the preparation and training of AI in distributed swarm-based systems, where local model-level errors may influence collective dynamics. This results in increased uncertainty during design, deployment, and operation, especially in tasks involving autonomous navigation and geospatial analysis. In such environments, perception errors may propagate across agents and amplify through collective swarm interactions.

This study develops a methodological framework for evaluating lifecycle risks of an AI-based Geospatial Perception Model (AGPM) for swarm UAV systems. The framework is designed for models responsible for recognizing and interpreting geospatial objects in support of autonomous decision-making.

The methodology is guided by three key requirements:

- It must adapt to the specific demands of autonomous navigation and swarm UAV operation.
- It must provide transparent and interpretable results that support engineering decisions.
- It must rely on established analytical tools capable of formalizing expert judgment into quantitative risk estimates.

The study is structured around the following research questions:

- Which risks, related to human and technological factors, are most significant during the preparation and training of AI models for swarm UAV systems?
- How can these risks be systematically identified, classified, and prioritized across the AGPM lifecycle?
- How can risk evaluation remain methodologically rigorous while still adapting to different operational and organizational contexts?
- What limitations do contemporary general-purpose multimodal AI systems exhibit when analyzing graphical representations of geospatial objects, and how do these limitations influence the risk profile of swarm UAV systems?

3. Background

Modern UAVs, including swarm-based UAV systems, are increasingly expected to operate autonomously in complex geospatial environments where satellite navigation may be unreliable or unavailable [5]. In such conditions, AI (particularly computer vision and machine learning) becomes essential for recognizing and interpreting environmental structures [6, 7].

Recent studies examine the use of convolutional neural networks, transformer-based architectures, and hybrid models for processing aerial imagery, satellite data, and onboard video streams from UAV sensors [8, 9]. These approaches support such tasks as localization, semantic segmentation, obstacle detection, landmark recognition, and environmental mapping [10]. Their performance, however, depends critically on the quality and representativeness of training data, as well as on the robustness of the training process [10, 11].

Navigation in dense urban areas, forests, mountainous terrain, and low-visibility conditions further increases system complexity. AI models must operate on noisy, incomplete, or heterogeneous input data, which elevates the probability of perception errors. Although many studies report improvements in accuracy metrics, less attention is given to the structured evaluation of risks emerging from dataset preparation, model configuration, and operational deployment.

In swarm-based systems, such risks are not isolated, because the multiagent structure amplifies local inconsistencies. The so-called “split-brain” problem (fragmentation of the swarm into multiple clusters) [12], together with perception errors or incorrect interpretation of geospatial objects by individual agents, can result in cumulative errors at the swarm level. Distributed multiagent systems are known to exhibit sensitivity to local perturbations, especially when coordination relies on shared perception or consensus mechanisms. Small inconsistencies in environmental interpretation may affect alignment between agents and lead to macroscopic deviations from intended collective behavior. Such dynamics are well documented in studies of emergent swarm coordination and consensus behavior [13–15], highlighting the importance of analyzing perception-related vulnerabilities from a systems perspective. Even minor deviations in training data or model configuration may trigger emergent behaviors that are difficult to trace back to their origin.

Another important issue concerns generalization [16]. Models trained on limited or synthetic datasets often perform well under controlled laboratory conditions but quickly degrade when exposed to variable lighting, weather, and complex spatial geometry [17, 18]. This gap between validation and real-world deployment introduces additional lifecycle risks.

Existing approaches to AI risk assessment are generally formulated at a broad level and do not explicitly address the specific challenges of preparing and training models for geospatial object recognition in multiagent systems [19]. Structured methodologies that allow systematic evaluation of risks related to training data quality, model training processes, and subsequent deployment within UAV swarm systems are largely limited in publicly available literature [16].

Particular attention should be paid to the problem of confabulation in the context of preparing and training AI systems for geospatial analysis – that is, the generation of plausible but factually incorrect or non-existent perceptual elements by a model [20, 21]. Although the term “confabulation” is most often associated with large language models, similar effects are observed in computer vision systems, including semantic segmentation, object detection [22], and spatial scene reconstruction [23]. Confabulations are especially dangerous in autonomous navigation tasks within complex geospatial environments, where the outputs of computer

vision systems directly influence trajectory planning, obstacle avoidance, and inter-agent coordination [24, 25]. Unlike random errors, confabulations tend to be systematic and may remain undetected during validation, since the generated internal representations appear logically consistent [20, 26].

Within swarm systems, these distortions can propagate across agents and produce collective misalignment. This increases the need for rigorous data preparation, controlled training strategies, and systematic evaluation of generalization behavior.

Taken together, existing research highlights the importance of a structured approach to AGPM lifecycle risk assessment in swarm-based UAV systems. Risks related to data quality, model generalization, perception reliability, and confabulation must be evaluated systematically at early stages of AI development.

4. Methodology

The proposed methodology addresses the absence of a structured framework for assessing risks associated with the preparation and training of AI in swarm-based UAV systems operating in complex geospatial environments. It relies on established analytical techniques and considers both human and technological factors that influence the reliability of AI-based perception and navigation.

The approach consists of three components.

1) Risk taxonomy development

a. Identified risks are grouped into two primary categories: human-related and technological factors.

b. The taxonomy covers all key stages of the AI lifecycle, including knowledge selection, knowledge analysis, model training, and model operation within swarm UAV systems.

2) Risk prioritization and evaluation

a. Expert judgment and pairwise comparisons are employed to assess the relative importance of individual risks across lifecycle stages.

b. The Analytic Hierarchy Process (AHP) is applied to quantify global risk priorities and to aggregate expert assessments into a consistent numerical framework.

3) Analysis of graphical images using AI

a. A series of experiments evaluates the ability of contemporary multimodal AI systems to analyze and interpret an aerial photograph of an urban scene representing a complex geospatial environment relevant to autonomous UAV navigation.

b. The stage examines whether such systems can reliably extract structured semantic information from aerial urban imagery (object identification, spatial relationships, scene composition, and contextual interpretation) rather than simply enumerate isolated visual elements.

c. The results provide empirical insights into the limitations in processing real-world geospatial imagery and help identify perception-related risks during AGPM preparation and training.

These components (risk taxonomy development, AHP-based prioritization, and empirical image analysis) form an integrated framework for evaluating AGPM readiness and associated risks in swarm UAV systems. The framework is not intended as a predictive statistical model; instead, it provides a structured basis for analyzing how risks emerge and evolve across the AI lifecycle.

4.1. Process of AI implementation

Deploying an AGPM within a swarm-based UAV system introduces risks linked to both technological and human factors across the stages of model preparation, training, and operation. The AGPM lifecycle can be represented by a structured model that illustrates how expert effort and computational resources are distributed throughout these stages.

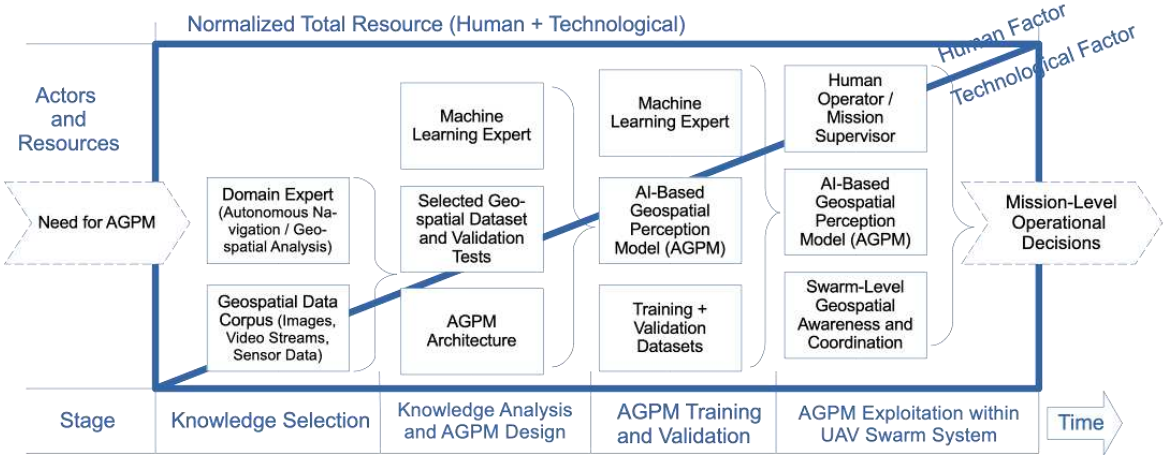


Figure 1. Conceptual model of resource allocation during the development and deployment of an AGPM in swarm-based UAV systems.

Figure 1 illustrates four stages of the AGPM lifecycle:

Knowledge Selection – this stage focuses on constructing a representative corpus of geospatial data. The dataset may include geospatial object drawings, aerial imagery, onboard UAV video streams, and supporting metadata required for accurate environmental interpretation. The output consists of curated training and testing datasets for subsequent model development and validation.

Knowledge Analysis – the assembled dataset is examined to determine relevant data subsets, define the neural network architecture, and select preprocessing and image representation techniques. The goal is to enable reliable interpretation of complex visual scenes, including environments with high object density, heterogeneous terrain, and varying illumination. The outcome is an AGPM architecture tailored to geospatial recognition tasks.

AGPM Training – performance criteria are defined, and the model undergoes iterative training and validation using the prepared datasets. This process may include dataset refinement, architectural adjustments, and hyperparameter tuning. Systematic recognition errors, reduced generalization, or geospatial confabulations may emerge here, introducing risks that can later affect swarm-level behavior.

Model Operation – this stage includes controlled testing of AGPM followed by deployment in real-world environments. The AGPM processes incoming visual and sensor data to support decision-making by individual swarm agents. Domain experts monitor performance and determine whether continued deployment is justified or model refinement is required in response to detected errors.

The stages of model modernization and decommissioning are not considered here, as the focus is on risks arising during preparation and training of AI models. Figure 1 outlines the boundaries of the model preparation process and illustrates the distribution of resources between human expertise and computational infrastructure. Early stages rely more heavily on expert involvement in dataset construction and evaluation design, whereas later stages shift

toward computational resource intensity. This distribution highlights the relationship between resource allocation and risk formation throughout the AGPM lifecycle.

4.2. Risks Classification

The potential risks associated with the preparation, training, and operation of AGPM for swarm-based UAV systems are summarized in Table 1.

The identified risks form a hierarchical structure in which each stage of the AGPM lifecycle – knowledge selection, knowledge analysis, model training, and model operation – is treated as a stage-specific source of risk. Within this hierarchy, human-related and technological factors are interpreted as primary risk categories that influence the reliability of geospatial perception, swarm-level situational awareness, and mission-oriented decision-making in complex environments.

Table 1. Risks and Mitigation Measures for AGPM Implementation.

Stage	Factor	Risk	Description	Risk Mitigation Measures
Knowledge selection	Human	Data representativeness	Training data do not sufficiently cover real-world landscapes, weather conditions, and operational scenarios relevant to swarm-based UAV systems.	Dataset expansion; data collection in diverse environments; expert validation of scenario coverage
	Human	Dataset bias	Presence of hidden spatial or sensor-related biases within the training data	Dataset balancing; feature distribution analysis; cross-regional validation
	Human	Temporal bias	Data collected within a limited time frame (e.g., seasonal effects or pre-change landscapes)	Temporal stratification; periodic dataset updates; synthetic scene modifications
	Human	Scale mismatch	Underestimation of altitude and camera angle variations affecting the perceived scale of geospatial objects during dataset construction	Multiscale dataset design; multiscale training strategies; hierarchical spatial feature extraction
	Human	Limited scene diversity	Training corpus dominated by homogeneous routes or imaging conditions, failing to reflect real-world variability	Targeted scenario expansion; data augmentation; use of simulation and synthetic environments
Knowledge analysis	Human	Annotation errors	Incorrect or ambiguous labeling of geospatial objects in the dataset	Double annotation; inter-annotator agreement checks; active learning strategies
	Technological	Feature leakage	Model reliance on spurious features (e.g., shadows, noise, sensor artifacts) rather than meaningful geospatial structures	Feature importance analysis; ablation studies; visualization of learned representations
	Human	Inadequate model architecture	Suboptimal architecture selection due to insufficient consideration of scene complexity or real-time operational constraints	Comparative testing of alternative architectures; inference time profiling on target platforms
	Technological	Incorrect spatial-geometric interpretation	Flattened (2D) perception of complex 3D environmental structures	Depth estimation techniques; multi-view geometry; integration of spatial models
	Human	Lack of uncertainty estimation	Absence of formal uncertainty assessment requirements at the model design stage	Incorporation of uncertainty estimation into system requirements; probabilistic neural networks; ensemble methods

AGPM Training	Technological	Overfitting	Excessive adaptation of the model to training data, reducing generalization capability	Regularization; cross-validation; early stopping strategies
	Human	Domain shift	Insufficient consideration of differences between training and real operational environments	Domain-aware dataset design; domain adaptation techniques; evaluation on atypical data
	Technological	Confabulations	Generation of plausible but non-existent geospatial objects or spatial structures	Control datasets; multi-view verification; sensor fusion strategies
	Technological	Confidence miscalibration	Mismatch between predicted confidence levels and actual recognition accuracy	Probability calibration; calibration curve analysis
	Technological	Catastrophic forgetting	Loss of previously acquired knowledge during incremental training	Replay buffers; continual learning approaches
	Human	Incomplete validation criteria	Evaluation limited to overall accuracy without considering safety-critical error cases	Safety-oriented performance metrics; scenario-based testing
Model operation	Technological	Sensor artifacts	Noise, degradation, or calibration drift of onboard sensors	Periodic recalibration; signal filtering; multisensor fusion
	Technological	Concept drift	Gradual changes in environmental characteristics over time	Continuous monitoring; periodic model retraining
	Technological	Cascading error propagation	Errors of a single agent influencing collective swarm behavior	Local validation mechanisms; limitation of individual agent influence within swarm coordination protocols
	Human	Loss of situational awareness	Human operator fails to detect AI-related errors at early stages	Confidence visualization interfaces; explainable AI mechanisms

The classification presented in Table 1 reflects the authors' perspective and is intended to provide a structured methodological baseline. It may be expanded or refined by domain experts depending on specific application scenarios and operational requirements.

The proposed classification accounts for both the temporal dynamics of risk manifestation across the AGPM lifecycle and the source of responsibility for the formation of preconditions leading to their emergence. This ensures consistency between the theoretical lifecycle model of AGPM, and the practical risk taxonomy presented in Table 1.

As mentioned above, the risks form a hierarchy in which each stage of AGPM preparation and operation can be interpreted as a stage-specific source of risk, while the underlying factors represent either technological or human-related risk categories. Given this hierarchy and with the aim of formalizing the evaluation procedure, it is appropriate to apply the AHP proposed by T. Saaty [27, 28]. The use of AHP is consistent with the objectives of this study and enables a quantitative interpretation of the relative significance of the identified risks.

4.3. Mathematical Formalization of Risk Distribution

A mathematical representation of risk distribution is introduced to formalize the implementation stages of the AI model.

Figure 1 distinguishes four stages of AGPM implementation: knowledge selection, knowledge analysis, model training, and model operation. Each stage is treated as an indivisible unit with normalized length equal to one. Within each stage, 0 corresponds to its initiation and 1 to full completion. This normalization expresses stage progression in relative terms, independent of absolute execution time.

Let $f(t)$ denote the risk distribution function defined over interval $D\big(f(t)\big), t \in [0; 4]$, corresponding to the four consecutive stages. For a given value t, the stage index is determined by $k = [t]$, where $[\cdot]$ denotes floor operation. The degree of completion within the current stage is given by $\delta = t - [t]$, which corresponds to the fractional part of t.

For example, a value $t = 2.75$ indicates that the second stage has been fully completed, and the third stage has reached 75% completion.

The modeling domain is represented as a unit-area rectangle bounded by the time axis, the energy (or resource) expenditure axis, the completion boundary of AI implementation, and the initiation boundary of knowledge selection. This normalization allows risk or energy expenditure to be interpreted in relative terms.

The risk distribution function $f(t)$, is defined on $D\big(f(t)\big), t \in [0; 4]$ with properties analogous to a probability density function: $0 \leq f(t) \leq 1$ and $0 \leq \int_{D(f(t))} f\,(t) \leq 1$.

Let $f_0(t)$ denote the expected risk distribution function and $f_1(t)$ its observed approximation. Define a binary function $d(a, b) \equiv \begin{cases} a - b, & \text{if } (a - b) \geq 0 \\ 0, & \text{if } (a - b) < 0 \end{cases}$ that characterizes deviations between expected and observed risks.

Values $d\big(f_0(t), f_1(t)\big) > 0$ indicates observed technological risks exceeding expectations, whereas $d\big(f_1(t), f_0(t)\big) < 0$ correspond to excess human-related (anthropogenic) risks. The case $d\big(f_0(t), f_1(t)\big) = d\big(f_1(t), f_0(t)\big) = 0$ implies $f_0(t) = f_1(t)$.

4.4. Analysis of Graphical Images Using AI

The analysis of graphical images forms an integral part of the methodology for assessing risks associated with the preparation and training of AGPM in swarm-based UAV systems. This component focuses on the ability of contemporary multimodal AI systems to process and interpret aerial photographs of urban environments, including scenes with dense development, infrastructure fragments, damaged structures, and other complex spatial configurations.

Several publicly accessible multimodal AI tools were used to evaluate these capabilities, including Google Gemini [29], OpenAI ChatGPT [30], Microsoft Edge Copilot [31], xAI Grok [32], and Perplexity AI [33]. These systems are not specialized computer vision models and are not optimized for autonomous navigation or AGPM construction. Instead, they serve as representative examples of widely available general-purpose AI tools that may be applied to geospatial imagery outside controlled laboratory conditions. This approach allows observation of practical model behavior rather than laboratory-level accuracy of specialized algorithms.

As a test case, we used an aerial photograph of an urban environment exhibiting signs of infrastructure damage (repository ref. [34]). The original image source is StockCake (Available online: https://stockcake.com/i/urban-devastation-scene_1096383_1039940, accessed on 23 February 2026); the exact copy used in the experiment is agpm/photo/aerial_photo_urban_damage.png. The image includes buildings with varying degrees of structural damage, road network elements, vehicles, smoke-affected areas, and spatial obstacles. The scene is characterized by high object density, heterogeneous spatial structure, and uncertain object boundaries, making it suitable for evaluating semantic recognition, spatial interpretation, and navigation-oriented reasoning relevant to autonomous flight tasks.

Each AI system received the same image along with structured natural-language instructions. The prompts included: (1) identification of navigation-relevant scene elements;

(2) classification of detected elements into static obstacles, hazardous zones, and conditionally safe areas; (3) analysis of spatial relationships; (4) proposal of a risk-minimizing UAV trajectory; and (5) explicit self-assessment of assumptions and uncertainties.

This structured interaction design enabled evaluation not only of object recognition performance but also of semantic generalization, spatial reasoning, and decision-oriented interpretation within the context of swarm-based UAV operations.

5. Results

The normalized domain introduced in Section 4.3 is interpreted geometrically as a unit-area rectangle representing the lifecycle stages of AGPM preparation and operation (Fig. 1). Within this framework, risk magnitude is expressed as the relative area of a corresponding subregion. A diagonal separating the lower-left and upper-right corners serves as a conditional boundary between human-related and technological risks. This symmetric partition reflects an initial assumption of equal prior weights for the two categories. Subsequent deviations from this balance are determined through expert evaluation using the AHP.

5.1. Risk Analysis

Following [27, 28], the pairwise comparison scale presented in Table 2 was adopted for the AHP procedure.

Table 2. Pairwise comparison scale for risk assessment using AHP.

Intensity	Definition	Interpretation
1	Equal importance	Both activities contribute equally to the objective
2	Weak or slight	
3	Moderate importance	Experience and judgment slightly favor one activity over another
4	Moderate plus	
5	Strong importance	Experience and judgment clearly favor one activity over another
6	Strong plus	
7	Very strong or demonstrated importance	One activity is strongly favored over another
8	Very-very strong	
9	Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation

Risks evaluation was performed using AHP through construction of pairwise comparison matrices at each level of the hierarchy (stage / factor / specific risk), denoted as $A = \{a_{ij}\}$, where $a_{ij} = \frac{1}{a_{ji}}, a_{ii} = 1$. For each matrix, the principal eigenvector $\omega_i = \frac{\left(\prod_{j=1}^{n} a_{ij}\right)^{1/n}}{\sum_{k=1}^{n}\left(\prod_{j=1}^{n} a_{kj}\right)^{1/n}}$ was computed to obtain local normalized weights. The maximum eigenvalue $\lambda_{max} = \frac{1}{n}\sum_{i=1}^{n}\frac{(A\omega)_i}{\omega_i}$, the consistency index $CI = \frac{\lambda_{max}-1}{n-1}$, and the consistency ratio $CR = \frac{CI}{RI}$ were also determined. Here, n denotes the dimension of the matrix, and RI represents the corresponding “random index”.

AHP weights and consistency checks were computed using our Python implementation (repository ref. [34], directory **ahp/**). To illustrate the procedure, selected pairwise comparison matrices and their associated consistency indices are presented below.

Table 3. Pairwise comparison matrix for the stages of AGPM implementation.

	Knowledge selection	Knowledge analysis	AGPM Training	Model Operation	Normalized Weight
Knowledge selection	1	1/2	1/3	1/3	0.10
Knowledge analysis	2	1	1/2	1/3	0.15
AGPM Training	3	2	1	1/4	0.23
Model Operation	3	3	4	1	0.52

[1] Consistency index: $CR = 0.08$

Representative examples include the pairwise comparison matrices for human-factor risks at the knowledge selection stage (Table 4), human-factor risks at the knowledge analysis stage (Table 5), and technological risks at the AGPM training stage (Table 6).

Table 4. Pairwise comparison of human-factor risks at the knowledge selection stage.

	$R_{1.1.1}$	$R_{1.1.2}$	$R_{1.1.3}$	$R_{1.1.4}$	$R_{1.1.5}$	Normalized Weight
$R_{1.1.1}$	1	1	2	2	3	0.30
$R_{1.1.2}$	1	1	2	2	3	0.30
$R_{1.1.3}$	1/2	1/2	1	1	1/2	0.12
$R_{1.1.4}$	1/2	1/2	1	1	1/2	0.12
$R_{1.1.5}$	1/3	1/3	2	2	1	0.16

[1] Consistency index: $CR = 0.04$.

here

- $R_{1.1.1}$ – data representativeness
- $R_{1.1.2}$ – dataset bias
- $R_{1.1.3}$ – temporal bias
- $R_{1.1.4}$ – scale mismatch
- $R_{1.1.5}$ – limited scene diversity

Table 5. Pairwise comparison of human-factor risks at the knowledge analysis stage.

	$R_{2.1.1}$	$R_{2.1.2}$	$R_{2.1.3}$	Normalized Weight
$R_{2.1.1}$	1	1/7	3	0.15
$R_{2.1.2}$	7	1	9	0.79
$R_{2.1.3}$	1/3	1/9	1	0.07

[1] Consistency index: $CR = 0.07$.

here

- $R_{2.2.1}$ – annotation errors
- $R_{2.2.2}$ – inadequate model architecture
- $R_{2.2.3}$ – lack of uncertainty estimation

Table 6. Pairwise comparison of technological risks at the AGPM training stage.

	$R_{3.2.1}$	$R_{3.2.2}$	$R_{3.2.3}$	$R_{3.2.4}$	Normalized Weight
$R_{3.2.1}$	1	1/4	2	1/5	0.10
$R_{3.2.2}$	4	1	5	1/3	0.28
$R_{3.2.3}$	1/2	1/5	1	1/7	0.06
$R_{3.2.4}$	5	3	7	1	0.56

[1] Consistency index: $CR = 0.04$.

here

- $R_{3.2.1}$ – overfitting
- $R_{3.2.2}$ – confabulations
- $R_{3.2.3}$ – confidence miscalibration
- $R_{3.2.4}$ – catastrophic forgetting

The standard AHP procedure typically involves a panel of domain experts and AI specialists. Domain experts should possess experience in swarm navigation and autonomous systems, while AI specialists should have expertise in areas such as computer vision, semantic segmentation, etc. Expert selection is based on demonstrated domain competence, familiarity with AI-assisted systems, and the absence of conflicts of interest.

In the present study, the coefficients used in the pairwise comparison matrices were determined by the authors based on domain expertise and structured analysis of inter-risk relationships. All matrices satisfied the consistency condition $CR < 0.1$, indicating acceptable internal coherence of the judgments.

Local normalized weights were obtained for risks at each hierarchical level. Global risk weights were computed by multiplying the corresponding stage weight, factor weight, and local risk weight:

$$\omega_{global} = \omega_{stage} \times \omega_{factor} \times \omega_{risk}$$

A consolidated ranking of risks based on global weights is presented in Table 7.

Table 7. Consolidated risk ranking based on global weights.

Stage / Factor / Risk	Global Weight
Model Operation / Technological / Cascading Error Propagation	0.306
Model Operation / Technological / Sensor Artifacts	0.102
Knowledge analysis / Human / Inadequate Model Architecture	0.099
AGPM training / Technological / Catastrophic Forgetting	0.096
Model Operation / Human / Loss of Situational Awareness	0.065
AGPM training / Technological / Confabulations	0.048
Model Operation / Technological / Concept Drift	0.046
AGPM training / Human / Incomplete Validation Criteria	0.038
Knowledge selection / Human / Data Representativeness	0.030
Knowledge selection / Human / Dataset Bias	0.030
AGPM training / Human / Domain Shift	0.019
Knowledge analysis / Human / Annotation Errors	0.019
Knowledge analysis / Technological / Feature Leakage	0.017
AGPM training / Technological / Overfitting	0.017
Knowledge selection / Human / Limited Scene Diversity	0.016
Knowledge selection / Human / Scale Mismatch	0.012
Knowledge selection / Human / Temporal Bias	0.012
AGPM training / Technological / Confidence Miscalibration	0.010
Knowledge analysis / Technological / Incorrect Spatial-Geometric Interpretation	0.008
Knowledge analysis / Human / Lack of Uncertainty Estimation	0.008

The AHP-based evaluation indicates that the operation stage of AGPM within swarm-based UAV systems contributes most strongly to the overall risk profile. During operation, local perception or data-processing errors made by individual agents can propagate through the swarm, resulting in cascading error amplification and undesirable collective behavior. At this

stage, such errors may also produce the most severe consequences for safety and mission performance.

Global weight analysis shows that cascading error propagation between agents represents the dominant mechanism of collective vulnerability in the analyzed hierarchy among all identified risks. This result reflects the inherent properties of multiagent systems, where minor perception inaccuracies may be amplified through collective interaction. Sensor-related risks also exhibit high significance, underscoring the influence of physical sensing limitations and environmental degradation on input data quality under operational conditions.

The results further emphasize the role of the human factor, particularly during the knowledge analysis stage. Inappropriate selection of AGPM architecture ranks among the highest-weighted risks, indicating the long-term impact of early architectural decisions on system robustness. Risks related to operator training and supervision of AGPM outputs likewise maintain measurable influence during operation.

Geospatial confabulations demonstrate a moderate but non-negligible global weight. Although less dominant than operational technological risks, their impact remains significant. Unlike random recognition errors, confabulations may generate internally coherent yet factually incorrect spatial representations. In autonomous navigation tasks, especially within swarm interaction dynamics, such distortions can lead to critical consequences, especially when amplified by swarm-level interaction mechanisms.

Overall, the evaluation reveals a significant asymmetry in lifecycle risk distribution. ***Early stages establish latent vulnerabilities, while the most severe consequences emerge during operation.*** Errors accumulated during knowledge selection, analysis, and training may later manifest as cascading failures or emergent swarm-level effects. These findings reinforce the need for comprehensive risk assessment that integrates both human and technological factors across the entire AI lifecycle.

5.2. Analysis of Graphical Images Using AI

The comparative analysis of responses (repository ref. [34], directory agpm/) generated by different multimodal AI systems shows a high level of consistency in semantic scene recognition (Fig. 2, Semantic Recognition). All tested tools correctly identified the general environment type, primary static objects, and potentially hazardous zones. This indicates reliable performance in tasks requiring general visual context understanding.

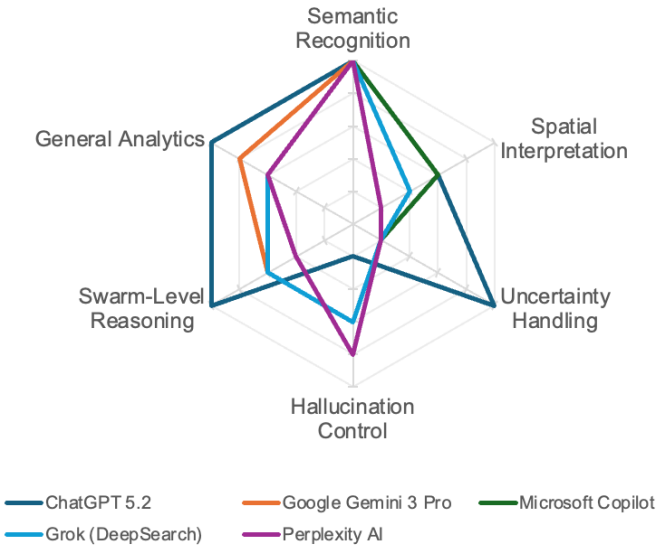


Figure 2. Comparative analysis of multimodal AI system performance in interpreting an aerial urban scene.

Despite this consistency, several systemic limitations emerged. Spatial interpretation remained largely heuristic and was not supported by metric reconstruction of scene geometry or formal reliability assessment. In most cases, the models did not explicitly quantify uncertainty, increasing the likelihood that inferred assumptions could be interpreted as verified conclusions. As reflected in the aggregated model profiles (Fig. 2), indicators related to spatial reasoning and uncertainty reflection display greater variability and lower stability than basic semantic recognition.

The models also exhibited a tendency to produce approximate numerical estimates without sufficient justification in the input data (so-called “pseudo-quantification”). This behavior corresponds to the previously identified risk of confabulation and may introduce additional hazards in autonomous navigation tasks. In Fig. 2, this tendency is visible in the uneven performance across criteria related to logical consistency and self-assessment.

The similarity of performance patterns across platforms suggests that these limitations are systemic rather than implementation-specific. The observed constraints likely reflect structural characteristics of the current generation of general-purpose multimodal AI systems when applied to precise spatial analysis tasks.

In swarm-based UAV systems, such limitations may be amplified through collective interaction mechanisms. Minor spatial interpretation errors can propagate between agents and lead to emergent effects at the swarm level.

These findings highlight the importance of a structured risk assessment methodology that accounts not only for technical performance metrics but also for behavioral characteristics of AI systems when processing geospatial information.

5.3. Collective Error Amplification Mechanism

The AHP-based results identified cascading error propagation as the dominant global risk during the operational stage of AGPM deployment. To interpret this result from the perspective of swarm dynamics, we formalize the mechanism through which local perception errors may evolve into collective deviations.

A swarm-based UAV system can be represented as a distributed dynamical system defined on an interaction graph $G = (V, E)$, where V denotes the set of agents and E represents communication or coordination links. Let the internal state of agent i at time t be denoted by $x_i(t)$. Environmental perception provided by the AGPM is modelled as

$$\hat{s}_i(t) = s(t) + \varepsilon_i(t)$$

where $s(t)$ is the true environmental signal and $\varepsilon_i(t)$ is the local perception error of agent i.

In many swarm coordination schemes (particularly consensus-based interaction models widely studied in distributed multi-agent systems [13]) the agent state update can be written [13, 35] as:

$$x_i(t+1) = x_i(t) + \alpha \sum_{j \in \mathcal{N}_i} \omega_{ij}(\mathrm{x}_j(t) - x_i(t)) + \beta \hat{s}_i(t),$$

where $\mathcal{N}_i$ denotes the neighborhood of agent i, ω_{ij} are interaction weights, $\alpha > 0$ is the coordination gain, and $\beta > 0$ reflects the influence of perceived environmental information.

Substituting the perception model yields

$$x_i(t+1) = x_i(t) + \alpha \sum_{j \in \mathcal{N}_i} \omega_{ij}(x_j(t) - x_i(t)) + \beta s(t) + \beta \varepsilon_i(t),$$

In compatc vector form, the collective dynamics become

$$x(t+1) = (I - \alpha L)x(t) + \beta s(t) + \beta \varepsilon(t),$$

where L is the graph Laplacian of the swarm interaction topology [13, 35, 36], $s(t)$ is the environmental input vector shared across agents, and $\varepsilon(t)$ is the vector of local perception errors.

This formulation highlights that perception errors do not remain confined to individual agents. Their system-level impact depends on:

- the coordination gain α,
- the algebraic connectivity $\lambda_2(L)$ of the interaction graph,
- the statistical properties and correlation structure of $\varepsilon(t)$.

Under sufficiently strong inter-agent coupling or highly connected topologies, correlated perception errors may propagate through the network before consensus stabilization occurs. The collective deviation from the nominal trajectory $x^*(t)$ (obtained when $\varepsilon(t) = 0$) can be conceptually bounded as

$$\|x(t) - x^*(t)\| \leq \gamma\big(\alpha, \lambda_2(L)\big)\|\varepsilon(t)\|,$$

where $\gamma(\cdot)$ represents an amplification factor determined by coordination strength and network structure; $\|\cdot\|$ denotes the Euclidean norm.

Provided that the coordination gain satisfies classical consensus stability conditions for linear distributed systems [13], the homogeneous dynamics remain stable. However, persistent or correlated perception perturbations may lead to bounded yet amplified collective deviations.

This representation is illustrative and intended to clarify the structural mechanism of error amplification rather than to provide a complete stability proof. It demonstrates that vulnerabilities introduced during AGPM preparation and training may act as persistent perturbations within distributed swarm dynamics. Consequently, systematic perception inaccuracies (particularly those arising from architectural bias, dataset imbalance, or confabulation effects) can manifest at the collective level as emergent coordination failures.

Thus, the dominance of cascading error propagation in the AHP ranking is consistent with the intrinsic sensitivity of distributed multi-agent systems to correlated perception disturbances.

6. Conclusions

The study examined how perception-level uncertainties introduced during the preparation and training of AI-based geospatial perception models (AGPM) influence the stability of swarm-based UAV systems. Instead of treating the risk purely as an engineering issue, the framework links lifecycle-related vulnerabilities to distributed multiagent dynamics. Qualitative risk classification, together with quantitative prioritization using the Analytical Hierarchy Process (AHP), helps clarify how such vulnerabilities emerge and propagate across development stages.

The results show that human-related factors play a decisive role during early AGPM development, particularly at the stages of knowledge selection and knowledge analysis. Architectural and dataset-related decisions made at these stages shape the structural conditions for later system behavior. Such vulnerabilities may remain latent and become visible only during deployment. In contrast, technological risks dominate during operation, with cascading error propagation among swarm agents receiving the highest

global weight. This reflects a basic property of distributed systems: local perception errors can spread through interaction and affect collective behavior.

The comparative analysis of contemporary multimodal AI systems further shows that, although semantic scene recognition is generally reliable, spatial-geometric reconstruction and formal uncertainty estimation remain limited. The observed limitations (including incorrect spatial interpretation, feature distortion, pseudo-quantification, and confabulation) correspond to the perception-related risks identified in the hierarchical model. This alignment supports the practical relevance of these risks for autonomous navigation in complex geospatial environments.

From a swarm intelligence perspective, collective robustness cannot be inferred from individual agent accuracy alone. Even moderate perception-level disturbances may trigger amplification effects under consensus-based or alignment-based coordination. Robustness of multiagent swarms should therefore be considered not only at the control or communication level, but also at the stage of perception-model preparation.

The proposed methodology is not intended to provide a complete catalog of risks. Its role is to offer a structured way of linking AI preparation processes with distributed swarm stability. By treating perception errors as perturbations within swarm dynamics, the study relates training-stage vulnerabilities to collective behavior. Future work should explore this relationship further through more detailed stability analysis and simulation-based validation in controlled multiagent environments.

Abbreviations

The following abbreviations are used in this manuscript:

AGPM	AI-based Geospatial Perception Models
AHP	Analytical Hierarchy Process
AI	Artificial Intelligence
UAV	Unmanned Aerial Vehicle

Appendix

Author Contributions: Oleksandr Kravchuk: methodology, software, validation, investigation, writing original draft, review and editing; Artem Taranowski: methodology, investigation, resources, writing original draft, review and editing; Dmytro Sinko: methodology, investigation, writing original draft; Hryhoriy Kravtsov: conceptualization, methodology, writing original draft, review and editing, supervision. Victor Samoylov: conceptualization, supervision.

Code Availability: The experimental code and artifacts used in this study are publicly available in a GitHub repository: https://github.com/oleksandrkravchukatpimee/swarm-ai-risks-assessment. The repository is organized into two modules: (i) **ahp** – Python scripts and outputs for the AHP-based risk evaluation; and (ii) **agpm** – image, prompts, and AI-generated analysis outputs for the graphical image experiment. A tagged release corresponding to this manuscript is provided in the repository.

Declarations

Conflict of interest: The authors declare no conflict of interest.
Ethical approval: Not applicable.
Consent to participate: Not applicable.

Consent for publication: Not applicable.

References

1. Brambilla M, Ferrante E, Birattari M, Dorigo M. (2013). Swarm robotics: a review from the swarm engineering perspective. Swarm Intell, 7:1–41. https://doi.org/10.1007/s11721-012-0075-2

2. Kravchuk OA, Samoilov VD. (2024). Overview of Swarm Systems' Anatomy. Interaction Problem. Èlektron Model, 46:101–121. https://doi.org/10.15407/emodel.46.02.101

3. Kravchuk OA, Samoilov VD. (2024). Application of Artificial Intelligence for Swarm Systems Managment. Èlektron Model, 46:29–42. https://doi.org/10.15407/emodel.46.06.029

4. National Institute of Standards and Technology (US). (2024). Artificial intelligence risk management framework : generative artificial intelligence profile. National Institute of Standards and Technology (U.S.), Gaithersburg, MD. https://doi.org/10.6028/NIST.AI.600-1

5. Chang Y, Cheng Y, Manzoor U, Murray J. (2023). A review of UAV autonomous navigation in GPS-denied environments. Robot Auton Syst, 170:104533. https://doi.org/10.1016/j.robot.2023.104533

6. Lu Y, Xue Z, Xia G-S, Zhang L. (2018). A survey on vision-based UAV navigation. Geo-Spat Inf Sci, 21:21–32. https://doi.org/10.1080/10095020.2017.1420509

7. Al-Kaff A, Martín D, García F, et al. (2018). Survey of computer vision algorithms and applications for unmanned aerial vehicles. Expert Syst Appl, 92:447–463. https://doi.org/10.1016/j.eswa.2017.09.033

8. Wang R, Ma L, He G, et al. (2024). Transformers for Remote Sensing: A Systematic Review and Analysis. Sensors, 24:3495. https://doi.org/10.3390/s24113495

9. Aleissaee AA, Kumar A, Anwer RM, et al. (2023). Transformers in Remote Sensing: A Survey. Remote Sens, 15:1860. https://doi.org/10.3390/rs15071860

10. Katkuri AVR, Madan H, Khatri N, et al. (2024). Autonomous UAV navigation using deep learning-based computer vision frameworks: A systematic literature review. Array, 23:100361. https://doi.org/10.1016/j.array.2024.100361

11. Wang Z. (2024). Sensitivities and where to find them: Domain shift robustness, attacks, and training variations in visual learning. Delft University of Technology. https://doi.org/10.4233/uuid:21e5b9c2-3e58-495c-881c-634e67ebe645

12. Sinko DP, Sinko KD. (2025). Analysis of the Applicability of Machine Learning Methods in Solving the Problem of Predicting the Implementation of Cluster Batching Factors. Èlektron Model, 47:22–39. https://doi.org/10.15407/emodel.47.01.022

13. Olfati-Saber R, Fax JA, Murray RM. (2007). Consensus and Cooperation in Networked Multi-Agent Systems. Proc IEEE, 95:215–233. https://doi.org/10.1109/JPROC.2006.887293

14. Zheng C, Lee K. (2023). Consensus decision-making in artificial swarms via entropy-based local negotiation and preference updating. Swarm Intell, 17:283–303. https://doi.org/10.1007/s11721-023-00226-3

15. Vicsek T, Czirók A, Ben-Jacob E, et al. (1995). Novel Type of Phase Transition in a System of Self-Driven Particles. Phys Rev Lett, 75:1226–1229. https://doi.org/10.1103/PhysRevLett.75.1226

16. Lyu S, Zhao Q, Zhou Z, et al. (2025). Deep Learning Based Domain Adaptation Methods in Remote Sensing: A Comprehensive Survey. https://doi.org/10.48550/arXiv.2510.15615

17. Cao K, Wang S, Wei Z, et al. (2024). Unsupervised Domain Adaptation Semantic Segmentation of Remote Sensing Imagery with Scene Covariance Alignment. Electronics, 13:5022. https://doi.org/10.3390/electronics13245022

18. Wang Y, Wen R, Ishii H, Ohya J. (2025). Multi-Weather DomainShifter: A Comprehensive Multi-Weather Transfer LLM Agent for Handling Domain Shift in Aerial Image Processing. J Imaging, 11:395. https://doi.org/10.3390/jimaging11110395

19. Tabassi E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology (U.S.), Gaithersburg, MD. https://doi.org/10.6028/NIST.AI.100-1

20. Sui P, Duede E, Wu S, So RJ. (2024). Confabulation: The Surprising Value of Large Language Model Hallucinations. https://doi.org/10.48550/arXiv.2406.04175

21. Huang L, Yu W, Ma W, et al. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans Inf Syst, 43:1–55. https://doi.org/10.1145/3703155

22. Kayhan OS, Vredebregt B, Gemert JC van. (2021). Hallucination In Object Detection -- A Study In Visual Part Verification. https://doi.org/10.48550/arXiv.2106.02523

23. Lu H, Su Y, Zhang X, et al. (2025). VisHall3D: Monocular Semantic Scene Completion from Reconstructing the Visible Regions to Hallucinating the Invisible Regions. https://doi.org/10.48550/arXiv.2507.19188

24. Nahavandi S, Alizadehsani R, Nahavandi D, et al. (2025). A Comprehensive Review on Autonomous Navigation. ACM Comput Surv, 57:1–67. https://doi.org/10.1145/3727642

25. Kondo K, Tewari CT, Tagliabue A, et al. (2025). PRIMER: Perception-Aware Robust Learning-based Multiagent Trajectory Planner. https://doi.org/10.48550/arXiv.2406.10060

26. Liu H, Xue W, Chen Y, et al. (2024). A Survey on Hallucination in Large Vision-Language Models. https://doi.org/10.48550/arXiv.2402.00253

27. Saaty TL. (2008). Decision making with the analytic hierarchy process. Int J Serv Sci, 1:83. https://doi.org/10.1504/IJSSCI.2008.017590

28. Saaty TL. (1990). How to make a decision: The analytic hierarchy process. Eur J Oper Res, 48:9–26. https://doi.org/10.1016/0377-2217(90)90057-I

29. Google. (n.d.). Gemini (product overview). https://gemini.google/about. Accessed 23 Feb 2026

30. OpenAI. (n.d.). ChatGPT (product overview / model access context). https://openai.com/chatgpt/overview. Accessed 23 Feb 2026

31. Microsoft. (n.d.). Copilot in Microsoft Edge (product overview). https://www.microsoft.com/edge/features/copilot. Accessed 23 Feb 2026

32. xAI. (n.d.). Grok (product overview). https://x.ai/grok. Accessed 23 Feb 2026

33. Perplexity. (n.d.). Perplexity AI (product overview). https://www.perplexity.ai. Accessed 23 Feb 2026

34. Kravchuk O. (2026). swarm-ai-risks-assessment: Experimental artifacts for “Risk Assessment of AI-Based Geospatial Perception Models in Swarm UAV Systems”. GitHub repository. https://github.com/oleksandrkravchukatpimee/swarm-ai-risks-assessment. Accessed 23 Feb 2026

35. (2010). Distributed Consensus in Multi-vehicle Cooperative Control: Theory and Applications (Ren, W. and Beard, R.W.; 2008) [Book Shelf]. IEEE Control Syst, 30:85–86. https://doi.org/10.1109/MCS.2010.936430

36. Mesbahi M, Egerstedt M. (2010). Graph theoretic methods in multiagent networks. Princeton University Press, Princeton