Ten Years of the Collective Perception Benchmark in Swarm Robotics: Achievements and Challenges

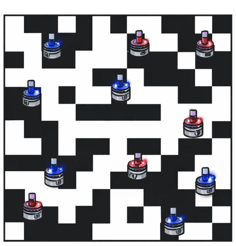
Heiko Hamann[1,2] and Andreagiovanni Reina[1,2,3]

[1] Centre for the Advanced Study of Collective Behaviour, Universität Konstanz, Konstanz, Germany
heiko.hamann@uni-konstanz.de, andreagiovanni.reina@gmail.com
[2] Department of Computer and Information Science, Universität Konstanz, Konstanz, Germany
[3] Department of Collective Behaviour, Max Planck Institute of Animal Behavior, Konstanz, Germany

Abstract. The collective perception scenario is an established swarm robotics task in which robots collectively infer a globally distributed environmental feature. The scenario is the most cited benchmark for collective decision-making in swarm robotics and has a 10-year history. Many variants of the original scenario have been proposed and studied. Many methods of collective decision-making have been tested against it. Given that the scenario was not initially intended as the defining benchmark, we summarize and analyze the literature and discuss the benefits and potential risks of a monoculture in benchmarking. We give a perspective on what could and should be improved in the future.

1 Introduction

Swarm robotics [19] generally lacks standardized benchmarks [23]. The collective perception (CP) scenario [57] is one of the few widely adopted testbeds and has been extensively used to evaluate methods that enable a swarm to reach consensus on the correct environmental state. In its original formulation, a swarm of robots operates on a grid where each tile (or grid cell) is either black or white (Fig. 1). The objective is to determine which color predominates across the entire environment. However, no individual robot has direct access to the global color



Generated with ChatGPT, OpenAI

Fig. 1: Illustration of the CP scenario. Robots locally sample tile color on a black-and-white grid, form and communicate opinions, and reach a consensus on the majority color. The benchmark was intentionally designed to avoid spatial correlations, allowing for locally unbiased sampling and motion-induced mixing that approximate a well-mixed population, to compare best-of-n strategies.

distribution as each can collect evidence only from the tile beneath it. Although CP generalizes to a best-of-n problem (i.e., selecting the best discrete option from a finite set of size n), most studies focus on the binary case ($n = 2$).

Conceptually, the scenario relates to the classical cellular automata majority problem posed by Mitchell et al. [28], which asked whether a decentralized system can determine if black or white cells are initially in the majority. It is also related to the collective estimation problem studied by Morlino et al. [29], where the real-valued density of black tiles is estimated rather than reduced to a binary decision. The CP scenario represents a deliberately simplified perception task that differs from richer perception problems, such as multi-robot SLAM [27], where robots build maps and localize themselves, and complex semantic perception [26], which requires interpreting object features. Instead, CP isolates a core challenge of swarm cognition: how a distributed system can reliably aggregate noisy, local information to infer a macroscopic variable; a process also observed in natural systems, such as insect colonies [29,43]. In applications, the macroscopic variable can be the presence or absence (or concentration) of a distributed resource (e.g., precious metals, pollutants), structural damage, or abnormal cell concentration (e.g., cancer cell detection). The CP scenario abstracts these problems into a model of distributed estimation and collective decision-making.

What the original authors of the CP scenario could not have anticipated ten years ago was its lasting influence. Its widespread adoption has improved comparability across collective decision-making methods, yet reliance on a single benchmark risks narrowing the design space. Although current approaches appear largely exploratory and show no clear signs of overfitting, the inflation of scenario variants and increasingly sophisticated algorithms complicates fair comparison. Over the past decade, CP has been extended from binary majority detection to density estimation, from uniform to spatially clustered environments, and from simple bio-inspired voting rules to Bayesian inference, evolutionary methods, reinforcement learning (RL), and blockchain-based robustness mechanisms. The complexity of communication models was expanded beyond 1-bit opinion exchange to include, for example, the statistical model parameters. While these novel methods advance the field, they also risk losing the benchmark's original simplicity. To move forward, we need a broader set of benchmarks that evaluate robustness, scalability, and security in a more systematic and comparable way.

We give an overview of the CP scenario, from its origin (Sec. 2) to various extensions in terms of the capabilities of the local perception units (i.e., the robots, Sec. 3), the environment to perceive (Sec. 4), and how to compare performance (Sec. 5). A meaningful benchmark requires not only a well-defined task and its variants (Sec. 4), but also explicit assumptions about the computational resources (robot capabilities, Sec. 3) to ensure a fair comparison. Moreover, the intended downstream use of information perceived by the swarm should inform the selection of models and evaluation metrics. We focus on three questions: which assumptions vary across studies, under which extensions prior conclusions no longer hold, and which metrics or generalizations remain underexplored.

2 Original Collective Perception Scenario

The original scenario by Valentini et al. [57] is formulated as a best-of-n problem with $n = 2$ options, that is, a binary collective decision-making problem. A swarm of robots operates in a bounded square arena whose floor is divided into a grid of black (b) and white (w) tiles, uniformly and randomly distributed in space (Fig. 1). The two colors correspond to the two decision options ($\mathcal{O} = \{b, w\}$), and their relative frequencies define the option quality (fill ratios ρ_b and $\rho_w = 1 - \rho_b$). The swarm's task is to collectively explore the environment and reach consensus (or a large majority) on the globally dominant option $o^* = \arg\max_{c \in \mathcal{O}} \rho_c$, that is, to determine whether $\rho_b > \rho_w$ or $\rho_b < \rho_w$. Robots cannot observe the global distribution ρ_b but can only take local measurements from ground sensors that detect the tile color beneath them. The CP scenario is meaningful only when individual estimations $\hat{\rho}_b^i$ are insufficiently reliable to determine the environmental state independently ($\mathbb{E}\left[(\hat{\rho}_b^i - \rho_b)^2\right] > 0$); otherwise, collaboration would neither improve the rate of correct perceptions of the true environmental state (i.e., perception accuracy) nor be required to reach an agreement. Because observations are strictly local and each robot samples only a small portion of the environment, individual estimates of the global distribution are often inaccurate. To enable a CP of the true environmental state, robots must exchange information. Each robot i maintains an opinion $o_i \in \{b, w\}$ about which color is more frequent and shares it locally through short-range communication. In the collective estimation variant, the swarm must agree on the real-valued density of black tiles, which defines the task as: $\min |\hat{\rho}_b^{\text{swarm}} - \rho_b|$ [29].

3 Robot Capabilities

Robot capabilities range from minimal machines that store, process, and exchange only a few bits of information to powerful robots capable of substantial computation. When comparing solutions, it is essential to consider the requirements each algorithm imposes on the robots. As robot platforms and algorithms vary widely, so do the capabilities assumed by different studies. In the following, we review existing work and outline research challenges and opportunities regarding how robot capabilities shape CP.

3.1 Computation, communication bandwidth, and memory

The original scenario [57] compares three algorithms with minimal robot requirements. Robots can store a single opinion o_i and its estimated quality $\hat{\rho}^i$, communicate their preferred opinion o_i to nearby neighbors (not its quality), and process $M \geq 1$ messages. The simplest tested algorithm, the weighted voter model [58], restricts interaction to $M = 1$ message. Other tested algorithms are variants of the majority rule requiring $M \geq 3$ [59]. Under the majority rule, robots adopt the opinion held by the majority of their neighbors, whereas using the voter model, they copy the opinion of a randomly selected neighbor.

Many subsequent studies aimed to improve efficiency, particularly concerning the speed–accuracy tradeoff (see Sec. 5). The assumed robot capabilities of each study constrained the proposed algorithms and shaped their design methods.

Bio-inspired voting algorithms. For very simple robots with minimal capabilities, effective solutions are bio-inspired. Zakir et al. [62,61] compare the weighted voter model with the cross-inhibition model, inspired by honeybee house-hunting. Cross-inhibition is identical to the weighted voter model, except that robots receiving neighbor votes for a different opinion transition to an uncommitted state. Cross-inhibition consistently outperforms the weighted voter model, producing more accurate, cohesive, stable, and faster decisions while retaining similar simplicity [61]. It also allows swarms to break symmetries and reach consensus in symmetric environments (i.e., options of equal quality), a setting in which the weighted voter model performs poorly.

Bayesian algorithms. The 'Bayes Bot' approach of Ebert et al. [15] has a strong statistical foundation. Each robot acts as a Bayesian estimator, modeling the unknown fill ratio ρ_b with a Beta distribution ($\rho_b \sim Beta(\alpha, \beta)$), and continuously updating its posterior using both newly collected local binary observations and those received from neighbors. The algorithm adjusts its decision time to task difficulty, making decisions quickly when evidence is clear and slowly when it is ambiguous, while maintaining accuracy comparable to fixed-time methods.

Bayesian methods typically offer improved performance but require robots to compute and transmit more information. Shan and Mostaghim [48,47] make a particularly valuable contribution by conducting a careful and fair benchmarking of Bayesian approaches against other algorithms, explicitly accounting for the communication bandwidth required by each method, an aspect often overlooked. They introduce and evaluate several variants of Distributed Bayesian Belief Sharing (DBBS) [45,46], in which robots maintain Bayesian likelihoods over all hypotheses and exchange probabilistic beliefs rather than single-option votes. DBBS achieves high accuracy and fast consensus even with clustered features or reduced communication, at the cost of higher required bandwidth.

Other Bayesian-based approaches include Abdelli et al. [1], Chiu et al. (belief with PSO tuning) [11], and Bartashevich and Sanaz (evidence theory) [6,7].

Reinforcement Learning and Evolutionary Approaches. The only RL approach we are aware of is that of Hussein et al. [20], in which robots learn behaviors through a sequence of state–action–reward modules, producing a CP strategy that outperforms human-designed algorithms.

Evolutionary swarm robotics has been explored more extensively. Almansoori et al. [2,3] evolve small continuous-time recurrent neural networks that map locally sensed tile color and neighbor communication to the robot's opinion. The evolved networks integrate perceptual and social information over time and outperform the weighted voter model.

Kaiser et al. [21] analyze how fitness functions shape evolved decision-making. By comparing task-independent (prediction-error minimization) and task-specific rewards, they find that the latter yield higher accuracy than the majority rule and the voter model, with decision times between the two.

Outlook. The original assumption of homogeneous robots and environments is limiting, and several studies have begun to relax it. A promising research direction is to leverage individual heterogeneity to improve group performance cost-effectively. If more sophisticated robots can achieve higher accuracy, speed, or cohesion than simpler ones, an important open question is whether similar performance can be reached by embedding only a small proportion of sophisticated robots within an otherwise simple swarm. Because simpler robots are typically cheaper to produce and operate, mixed swarms may offer practical and resource-efficient solutions. Early results suggest that heterogeneous swarms for CP are indeed a promising direction [64].

3.2 Communication network

CP relies critically on information exchange among robots and, in general, the communication network topology can strongly influence collective decision-making dynamics [42]. In swarm robotics, communication is typically limited to interactions among neighbors. In the original study [57], robots move randomly, promoting spatial mixing of opinions and approximating a well-mixed state that minimizes spatial correlations between neighboring robots. Motion patterns and communication range jointly determine the communication topology.

Aust et al. [5] analyze the impact of communication range. A key counter-intuitive result is a less-is-more effect: with large communication ranges, the swarm loses the ability to integrate new evidence after option qualities change. High connectivity allows a majority favoring an outdated option to suppress minority discoveries before they spread. Reducing the communication range mitigates these lock-in effects.

Outlook. Future comparisons should incorporate network connectivity as a core factor. Such insights could inform decentralized control mechanisms that adapt communication range to improve performance (see metrics in Sec. 5).

3.3 Environmental sensing

Robots sense only the color beneath them and estimate option qualities from local observations. In the original approach [57], all robots experience the same sensing noise; however, real-world deployments are unlikely to be so uniform [38]. Heterogeneity may be designed (e.g., by mixing robots of different sophistication or sensor types) or unintentionally through wear, manufacturing tolerances, calibration errors, or software and hardware faults.

Chin et al. [8,9,10] study a collective estimation variant of the problem in which robots have different, initially unknown, noise levels. By enabling robots to estimate both their own and other robots' noise levels, they can collectively weigh opinions inversely proportional to expected noise. Compared to approaches that ignore these differences, their method yields improved performance.

Outlook. A promising research track is that individuals may weigh option attributes differently, resulting in conflicting preferences within the swarm. Similarly, in nature, some ants prioritize darkness while others favor entrance size or number when evaluating nest sites [17]. Such settings can be modeled as multi-objective decision problems. Future work should extend the CP benchmark to examine how groups can resolve individual conflicts while still reaching consensus. This is relevant, for example, to robot teams in industrial applications when robots can be owned or configured by different stakeholders, potentially operating with differing objectives or incentives [60]. A benchmark could incorporate additional decision dimensions (e.g., ambient light or humidity, in addition to tile color) to study multi-objective CP.

4 Complex environments

The original benchmark (Sec. 2) is simple and elegant, enabling clean comparisons between algorithms. Yet, advancing toward real-world applications requires consideration of more complex and realistic conditions.

4.1 Beyond binary decisions, $n > 2$ options, and continuous options

Originally, the CP scenario focused on binary decisions ($n = 2$). This simplifies the analysis (e.g., the two option qualities can be expressed as a single ratio) and assumes that, when more options are present, the dynamics are dominated by competition between the two strongest alternatives. However, theoretical work indicates qualitative changes in the dynamics when $n > 2$ [41,13]. For $n > 2$, the environment can be characterized by fill ratios $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$, the CP task is then defined as a functional $f(\boldsymbol{\rho})$, and different objectives become possible. In particular, one may distinguish (a) best-of-n selection, identifying $o^* = \arg\max_j \rho_j$, (b) majority detection, determining whether $\exists j : \rho_j > 1/2$, or (c) full distribution estimation, reconstructing $\hat{\boldsymbol{\rho}} \approx \boldsymbol{\rho}$. Mostaghim and colleagues [7,48] compared several methods for best-of-n and full distribution estimation with up to $n = 10$ in simulations. While some methods scaled well, others exhibited substantial performance degradation as n increased, indicating the importance of studying $n > 2$.

Outlook. It remains an open question how to design a benchmark with $n \gg 2$ that remains elegant, analytically tractable, and focused on the most relevant conditions. A natural qualitative extension is going to continuous decision-making (i.e., collective estimation), where robots estimate real-valued quantities. For example, the swarm needs to find consensus on the mean $\frac{1}{|\Omega|}\int_\Omega g(x)\,dx$ of a scalar field $g : \Omega \to \mathbb{R}$ within a bounded area $\Omega \subset \mathbb{R}^d$ [37]. This shift to collective estimation better matches realistic swarm-sensing tasks (e.g., mean pollutant concentration, temperature, or radiation level over an area) but also requires different methods, as categorical and continuous decisions pose fundamentally different challenges. One should also consider an early CP formulation [26] comparable to swarm SLAM [27], in which robots collaboratively scan and characterize objects to infer properties such as shape and function.

4.2 Spatial heterogeneity

In the original scenario, black and white tiles are uniformly distributed, so any location provides a representative estimate of the fill ratio (Fig. 1).

Bartashevich and Mostaghim [6] introduce non-uniform tile color distributions and categorize them by clustering geometry (e.g., clusters, patches, strips). Whereas task difficulty in the original benchmark depends only on the fill ratio, they show that difficulty is also determined by how features are spatially arranged. The most challenging setting is the highly heterogeneous "half–half" environment (one side predominantly black, the other predominantly white).

Kelly et al. [24] study another form of spatial heterogeneity where communication quality varies across the environment. Some regions partially degrade messages, while others block communication entirely. Robots do not know where these communication-denied areas are or how large they are.

Outlook. Future work should extend the benchmark by incorporating a model of spatial correlation, enabling a systematic study of how tile distributions affect algorithmic performance while capturing patterns relevant to real-world environments. It would also be valuable to relate this scenario to best-of-n problems with spatially segregated options (e.g., site selection [35,56,58]). Analyses of spatial heterogeneity should further account for robot motion patterns and how these shape individual quality estimates and correlations among neighbors' observations.

4.3 Temporal heterogeneity

In a static environment, a single robot could, in principle, solve CP optimally given enough time and memory to explore the entire space. In dynamic environments, however, swarms can detect changes and infer the state of the world far more quickly and accurately than any individual robot. Many decision-making algorithms nevertheless lock the swarm into a fixed consensus that cannot adapt to new evidence, as strong social reinforcement prevents revising earlier decisions. Studying time-varying environments is therefore essential for assessing robustness to reveal otherwise hidden limitations. Extensions to the benchmark introduce abrupt changes in option quality during runtime [33,34,15,5,64], enabling evaluation of swarm performance in dynamic environments where the best option suddenly switches.

Outlook. Only a few studies address dynamic environments, and the field still lacks a systematic treatment of time-varying conditions. Another important but largely unexplored aspect is the appearance and disappearance of options over time, as considered in other best-of-n scenarios [56]. Existing CP studies assume that robots know the option set $\mathcal{O}$ (e.g., black and white), but not the associated qualities. When the option set is unknown or changes over time, robots must first discover the available alternatives before estimating their qualities [25]. Research shows that adding independent option discovery causes several algorithms to fail, whereas cross-inhibition consistently reaches consensus quickly [62,61]. Incorporating discovery can be crucial for practical deployment, as real environments may feature dynamic and initially unknown options.

4.4 Byzantine faults

A Byzantine fault occurs when a system component behaves arbitrarily or maliciously, sending inconsistent or deceptive information to others. Swarm robots must be robust to such faults. Therefore, several studies have investigated variants of the CP scenario that incorporate different adversarial conditions.

Malfunctioning and Malicious Robots. Robots that deviate from their specified behavior are typically referred to as Byzantine robots [14,54]. Such deviations may result from malicious attacks or from ordinary malfunctions due to wear, errors, or damage. Extensions of the CP benchmark, therefore, consider scenarios in which a fraction of the swarm is Byzantine.

Research has studied robustness against simple forms of Byzantine behavior, such as stubborn robots (also called zealots) [61,53,54,55]. Stubborn robots permanently hold a fixed opinion, ignore all social information, and continuously broadcast their chosen option. By never updating their state, they can inject a persistent bias into the swarm's information flow. Depending on their proportion and distribution, they can trap the swarm in a decision deadlock or push it toward the collective selection of an inferior option [61].

Two main strategies have emerged: (a) designing algorithms that are inherently robust to misbehaving agents and (b) developing mechanisms to detect and neutralize Byzantine robots. For (a), simple bio-inspired algorithms, such as cross-inhibition, have been shown to outperform the original solutions in the presence of zealots [61]. For (b), several works integrate methods for identifying faulty agents. A notable approach uses blockchain technology to maintain a shared reputation system and exclude low-reputation robots from the collective decision [39]. Early applications of blockchains in swarm robotics were demonstrated in CP [53] and collective estimation [54,55] scenarios with zealots, where smart contracts record votes, validate information, and automatically detect and blacklist inconsistent or malicious agents.

Miscommunication. Zakir et al. [63] study communication noise in cross-inhibition by including message corruption, where exchanged votes may be flipped. Counterintuitively, moderate noise can improve swarm accuracy by preventing premature convergence to suboptimal consensus, an effect confirmed through models, simulations, and 50-robot experiments. Excessive noise, however, prevents consensus, making miscommunication a tunable factor governing the speed-accuracy tradeoff.

Outlook. Ensuring robustness to adversarial or malfunctioning robots is essential for real-world deployment. While initial work addresses simple Byzantine behaviors, future research should tackle more sophisticated, potentially colluding attackers capable of steering consensus toward suboptimal outcomes.

5 Collective perception metrics

Comparison of alternative methods can be based on several alternative metrics.

5.1 Speed-accuracy tradeoffs and beyond

Most studies focus on accuracy, the probability that the swarm reaches a consensus favoring the correct option. Accuracy is typically examined together with decision speed, the time required to reach consensus, revealing the speed-accuracy tradeoff: a swarm can only gain accuracy at the cost of time and vice versa. Less explored but equally important metrics include group cohesion [62,61], robustness to internal noise (e.g., sensing noise [9,10]) and external disturbances (e.g., zealots [61]), and adaptability to environmental changes [5,33,64].

Outlook. Relying solely on the speed-accuracy tradeoff may be too narrow, as it may conceal important strengths or weaknesses of algorithms. While accuracy offers a convenient binary assessment (correct/incorrect), complementary metrics, such as decision value (quality ρ of the chosen option) or decision regret (difference between the chosen option quality and the best quality), can provide better insights. Such metrics could reveal collective dynamics, such as value-sensitive responses [40,41], not yet examined in the CP benchmark.

5.2 Collective perception as part of other processes

The metrics discussed above evaluate CP in isolation, but in real applications, it would be embedded in a collective perception-action loop. Any perception would exist to initiate action [25].

Task allocation. CP may function as one component of a more complex collective behavior. Ebert et al. [16] propose an algorithm enabling swarms to self-allocate dynamically across three independent perception tasks of differing difficulty. Instead, Fuady et al. [18] and Atasoy Bingöl et al. [4] study a task-allocation strategy to resize the robot group assigned to each perception task, increasing the number of robots for difficult tasks to boost accuracy and reducing it for easier tasks to save resources.

Localization and hierarchical tasks. In spatially heterogeneous environments, robots may require relative localization to interpret local observations correctly, coupling localization with perception. Soorati et al. [51] explicitly incorporate this requirement. Instead, Soma et al. [50] treat CP as an evidence-gathering layer in a best-of-2 decision problem, using it to estimate the qualities of two sites before making a site-selection decision based on those estimates.

Outlook. Although a few studies embed CP within richer collective behaviors, none extends the benchmark to allow the swarm to recognize when the perception task is complete. As Khaluf et al. [25] note, reaching a decision is not enough: a swarm must also detect that a decision has been reached so it can act on it. This decentralized awareness problem corresponds to quorum sensing, a mechanism observed in nature (e.g., social insects [36]) and explored in other swarm robotics contexts [32,12,30]. Quorum sensing requires defining what constitutes a collective decision: What threshold qualifies as a quorum? How long must it persist to avoid mistaking random fluctuations for genuine consensus? Incorporating quorum sensing into the benchmark highlights the importance of cohesion and stability. For example, voter-based models [57] typically yield only majorities due to stochastic fluctuations, whereas cross-inhibition [61] yields stable, cohesive majorities that could support reliable quorum detection. Future research should therefore couple perception with subsequent collective actions, rather than treating perception in isolation.

5.3 Hardware testing

The CP benchmark has not yet been directly linked to a real-world application. Designed as a concise, non-spatial best-of-n problem, it favors analytical clarity over realism. A clear pathway to deployment has yet to be established. Future extensions will likely need either to move toward estimating more realistic features or to place stronger emphasis on hardware constraints. We next discuss work that takes the initial steps in this direction.

Despite the cost of physical experiments, many studies (including the original [57]) report results with real robots, for example, see [16,31,22,63,10]. Hardware experiments often reveal practical issues, such as unforeseen collective dynamics resulting from limited mixing or unreliable asynchronous communication.

One of the most sophisticated implementations is by Siemensma and Haghighat [49]. They realize the binary feature of the benchmark using vibrating and non-vibrating tiles on a metallic, actuated surface, with a swarm of up to ten miniature IMU-based vibration-sensing robots. Vibrating tiles are driven by stacked micro-motors, and because vibration propagates across the structure, robots must threshold and filter IMU readings to classify tile states. This setup represents a significant step toward real-world deployments and applications.

Outlook. Work on the CP benchmark has not yet translated into realistic application scenarios. The vibration-based system of Siemensma and Haghighat [49] represents a promising step toward structural health monitoring. Other potential applications lie in domains with severe sensing or communication constraints, such as underwater environments, as well as in agriculture or nanorobotics, where swarms may need to detect whether a distributed condition (e.g., pest presence or molecular concentration) exceeds a critical threshold before collectively triggering an autonomous intervention [44,52].

6 Discussion and Conclusion

Over the past decade, the CP scenario has served as a valuable reference task for studying collective decision-making in swarm robotics. Overall, the reviewed literature reflects a broad, exploratory effort along three main axes: modifying the capabilities of the robots (e.g., computation, memory, sensing, and communication bandwidth; Sec. 3), extending the perception task itself (e.g., spatially and temporally heterogeneous environments, adversarial settings, and multi-option or continuous problems; Sec. 4), and broadening the objectives and evaluation criteria beyond the classical speed-accuracy tradeoff (Sec. 5).

From this review, we draw three main conclusions. First, as expected, performance tends to improve with increased robot capabilities (e.g., richer internal state representations and more informative communication). Second, conclusions established under restrictive assumptions (homogeneous robots, static environments, and fully cooperative interactions) do not necessarily transfer well to more realistic conditions, such as spatial heterogeneity, dynamic environments, or adversarial agents. Third, there is no universally best approach: the relative ranking of algorithms depends strongly on the chosen objective and evaluation metrics (e.g., accuracy, decision time, robustness, cohesion, or adaptability), and methods optimized for one criterion can be suboptimal for another.

The shared CP scenario has helped compare methods and collectively advance the field by serving as a common reference task. However, we see two opposing risks: focusing too much on a single, highly abstract task may narrow the scope of studies, while the inflation of loosely defined variants without an overarching framework can limit meaningful comparisons. This creates a fundamental tension between balancing coherence and controlled diversity.

We argue for a more rigorous and structured benchmarking framework. Rather than treating CP as a single fixed scenario or as an unstructured collection of variants, it could be formalized as a benchmark family in which key assumptions are explicitly parameterized, including computational and memory resources, communication bandwidth and topology, spatial and temporal correlations, and the presence of adversarial agents. A systematic exploration of these dimensions would clarify which conclusions are robust and which are contingent on specific modeling choices. Initial steps toward unification already exist. For example, Zakir et al. [61] show that independent discovery, stubborn agents, and message corruption can be formulated as mathematically equivalent processes within a unified model. This formal equivalence helps consolidate seemingly distinct benchmark variants under a common analytical framework.

Future benchmarks should be guided by application-oriented realism to ensure that swarm robotics, as an engineering discipline, is oriented toward deployment. CP is too often studied in isolation, whereas applications require integrated perception–decision–action loops [25]. Our field would benefit from embedding CP within broader collective processes, linking it to the actions it informs and to other coordination mechanisms such as task allocation [16]. Initial work on hierarchical decision-making [50] and dynamic task allocation [18] offers promising starting points, but a systematic integration has not yet been initiated.

A community-driven benchmarking effort—with shared task definitions, parameter ranges, evaluation metrics, and reference implementations—could substantially improve reproducibility and comparability. Such an initiative would not replace CP, but situate it within a structured ecosystem of related benchmarks, balancing simplicity, generality, and realism. Swarm robotics must move beyond counting colors on a grid and toward enabling swarms to sense, interpret, and act within the complex, dynamic environments of real-world applications.

Acknowledgments. We acknowledge support from DFG Excellence Strategy – EXC 2117-422037984. We thank the anonymous reviewers for many helpful comments.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Abdelli, A., Yachir, A., Amamra, A., Khaldi, B.: Maximum likelihood estimate sharing for collective perception in static environments for swarm robotics. Robotica **41**(9), 2754–2773 (2023)
2. Almansoori, A., Alkilabi, M., Tuci, E.: A comparative study on decision making mechanisms in a simulated swarm of robots. In: IEEE Congress on Evolutionary Computation (CEC). pp. 1–8. IEEE (2022)
3. Almansoori, A., Alkilabi, M., Tuci, E.: On the evolution of adaptable and scalable mechanisms for collective decision-making in a swarm of robots. Swarm Intelligence **18**(1), 79–99 (2024)
4. Atasoy Bingöl, S., Töpfer, T., Kosub, S., Hamann, H., Reina, A.: Optimal scalability-aware allocation of swarm robots: From linear to retrograde performance via marginal gains. IEEE Transactions on Systems Man and Cybernetics: Systems **56**(in press) (2026)
5. Aust, T., Talamali, M.S., Dorigo, M., Hamann, H., Reina, A.: The hidden benefits of limited communication and slow sensing in collective monitoring of dynamic environments. In: International Conference on Swarm Intelligence (ANTS), LNCS, vol. 13491, pp. 234–247. Springer, Cham (2022)
6. Bartashevich, P., Mostaghim, S.: Benchmarking collective perception: New task difficulty metrics for collective decision-making. In: EPIA Conference on Artificial Intelligence. LNCS, vol. 11804, pp. 699–711. Springer, Cham (2019)
7. Bartashevich, P., Mostaghim, S.: Multi-featured collective perception with Evidence Theory: tackling spatial correlations. Swarm Intelligence **15**(1), 83–110 (2021)
8. Chin, K.Y., Khaluf, Y., Pinciroli, C.: Minimalistic collective perception with imperfect sensors. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8862–8868. IEEE (2023)
9. Chin, K.Y., Pinciroli, C.: Adaptive self-calibration for minimalistic collective perception by imperfect robot swarms. arXiv:2410.21546 (2024)
10. Chin, K.Y., Pinciroli, C.: BayesCPF: Enabling collective perception in robot swarms with degrading sensors. arXiv:2504.04774 (2025)
11. Chiu, D., Nagpal, R., Haghighat, B.: Optimization and evaluation of a multi robot surface inspection task through particle swarm optimization. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 8996–9002. IEEE (2024)
12. Cody, J.R., Adams, J.A.: An evaluation of quorum sensing mechanisms in collective value-sensitive site selection. In: International Symposium on Multi-Robot and Multi-Agent Systems (MRS). pp. 40–47. IEEE (2017)
13. Crosscombe, M., Lawry, J.: Collective preference learning in the best-of-n problem. Swarm Intelligence **15**(1-2), 145–170 (2021)
14. Dorigo, M., Pacheco, A., Reina, A., Strobel, V.: Blockchain technology for mobile multi-robot systems. Nature Reviews Electrical Engineering **1**(4), 264–274 (2024)
15. Ebert, J.T., Gauci, M., Mallmann-Trenn, F., Nagpal, R.: Bayes Bots: collective Bayesian decision-making in decentralized robot swarms. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 7186–7192. IEEE (2020)
16. Ebert, J.T., Gauci, M., Nagpal, R.: Multi-feature collective decision making in robot swarms. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). pp. 1711–1719. IFAAMAS (2018)

17. Franks, N.R., Dornhaus, A., Metherell, B.G., Nelson, T.R., Lanfear, S.a.J., Symes, W.S.: Not everything that counts can be counted: ants use multiple metrics for a single nest trait. Proceedings of the Royal Society B: Biological Sciences **273**(1583), 165–169 (2006)
18. Fuady, S., Tarapore, D., Soorati, M.D.: SubCDM: Collective decision-making with a swarm subset. In: IEEE International Conference on Intelligent Robots and Systems (IROS). pp. 2633–2638 (2025)
19. Hamann, H.: Swarm robotics: A formal approach. Springer, Cham (2018)
20. Hussein, A., Elsawah, S., Petraki, E., Abbass, H.A.: A machine education approach to swarm decision-making in best-of-n problems. Swarm Intelligence **16**(1), 59–90 (2022)
21. Kaiser, T.K., Potten, T., Hamann, H.: Evolution of collective decision-making mechanisms for collective perception. In: IEEE Congress on Evolutionary Computation (CEC). pp. 1–8. IEEE (2023)
22. Karagüzel, T.A., Turgut, A.E., Eiben, A., Ferrante, E.: Collective gradient perception with a flying robot swarm. Swarm Intelligence **17**(1), 117–146 (2023)
23. Kegeleirs, M., Birattari, M.: Towards applied swarm robotics: current limitations and enablers. Frontiers in Robotics and AI **12** (2025)
24. Kelly, T.G., Soorati, M.D., Zauner, K.P., Ramchurn, S.D., Tarapore, D.: Collective decision making in communication-constrained environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7266–7271. IEEE (2022)
25. Khaluf, Y., Simoens, P., Hamann, H.: The neglected pieces of designing collective decision-making processes. Frontiers in Robotics and AI **6** (2019)
26. Kornienko, S., Kornienko, O., Constantinescu, C., Pradier, M., Levi, P.: Cognitive micro-agents: individual and collective perception in microrobotic swarm. In: Proceedings of the IJCAI-05 Workshop on Agents in Real-Time and Dynamic Environments. pp. 33–42 (2005)
27. Lajoie, P.Y., Beltrame, G.: Swarm-SLAM: Sparse decentralized collaborative simultaneous localization and mapping framework for multi-robot systems. IEEE Robotics and Automation Letters **9**(1), 475–482 (2023)
28. Mitchell, M., Crutchfield, J.P., Hraber, P.T.: Evolving cellular automata to perform computations: mechanisms and impediments. Physica D: Nonlinear Phenomena **75**(1), 361–391 (1994)
29. Morlino, G., Trianni, V., Tuci, E.: Evolution of collective perception in a group of autonomous robots. In: International Joint Conference on Computational Intelligence (IJCCI). SCI, vol. 399, pp. 67–80. Springer, Heidelberg (2010)
30. Oddi, F., Reina, A., Trianni, V.: Minimalist protocols for quorum sensing in robot swarms. In: International Conference on Swarm Intelligence (ANTS), LNCS, vol. 14987, pp. 141–154. Springer, Cham (2024)
31. Pacheco, A., Strobel, V., Dorigo, M.: A blockchain-controlled physical robot swarm communicating via an ad-hoc network. In: International Conference on Swarm Intelligence (ANTS). LNCS, vol. 12421, pp. 3–15. Springer, Cham (2020)
32. Parker, C.A.C., Zhang, H.: Collective unary decision-making by decentralized multiple-robot systems applied to the task-sequencing problem. Swarm Intelligence **4**(3), 199–220 (2010)
33. Pfister, K., Hamann, H.: Collective decision-making with Bayesian robots in dynamic environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7245–7250. IEEE (2022)

34. Pfister, K., Hamann, H.: Collective decision-making and change detection with Bayesian robots in dynamic environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8814–8819. IEEE (2023)
35. Prasetyo, J., De Masi, G., Ferrante, E.: Collective decision making in dynamic environments. Swarm Intelligence **13**(3-4), 217–243 (2019)
36. Pratt, S.C.: Quorum sensing by encounter rates in the ant *Temnothorax albipennis*. Behavioral Ecology **16**(2), 488–496 (2005)
37. Raoufi, M., Hamann, H., Romanczuk, P.: Speed-vs-accuracy tradeoff in collective estimation: An adaptive exploration-exploitation case. In: International Symposium on Multi-Robot and Multi-Agent Systems (MRS). pp. 47–55. IEEE (2021)
38. Raoufi, M., Romanczuk, P., Hamann, H.: Individuality in swarm robots with the case study of Kilobots: Noise, bug, or feature? In: Proceedings of the Artificial Life Conference (ALIFE). p. 35. MIT Press (2023)
39. Reina, A.: Robot teams stay safe with blockchains. Nature Machine Intelligence **2**, 240–241 (2020)
40. Reina, A., Bose, T., Trianni, V., Marshall, J.A.R.: Effects of spatiality on value-sensitive decisions made by robot swarms. In: Distributed Autonomous Robotic Systems (DARS 2016): The 13th International Symposium, SPAR, vol. 6, pp. 461–473. Springer, Cham (2018)
41. Reina, A., Marshall, J.A.R., Trianni, V., Bose, T.: Model of the best-of-N nest-site selection process in honeybees. Physical Review E **95**(5), 052411 (2017)
42. Reina, A., Njougouo, T., Tuci, E., Carletti, T.: Speed-accuracy trade-offs in best-of-n collective decision making through heterogeneous mean-field modeling. Physical Review E **109**(5), 054307 (2024)
43. Schmickl, T., Möslinger, C., Crailsheim, K.: Collective perception in a robot swarm. In: International Workshop on Swarm Robotics (SR 2006). LNTCS, vol. 4433, pp. 144–157. Springer, Heidelberg (2007)
44. Schranz, M., Umlauft, M., Sende, M., Elmenreich, W.: Swarm robotic behaviors and current applications. Frontiers in Robotics and AI **7** (2020)
45. Shan, Q., Mostaghim, S.: Collective decision making in swarm robotics with distributed Bayesian hypothesis testing. In: International Conference on Swarm Intelligence. pp. 55–67. Springer (2020)
46. Shan, Q., Mostaghim, S.: Discrete collective estimation in swarm robotics with distributed Bayesian belief sharing. Swarm Intelligence **15**(4), 377–402 (2021)
47. Shan, Q., Mostaghim, S.: Benchmarking performances of collective decision-making strategies with respect to communication bandwidths in discrete collective estimation. In: International Conference on Swarm Intelligence (ANTS), LNCS, vol. 13491, pp. 54–65. Springer, Cham (2022)
48. Shan, Q., Mostaghim, S.: Many-option collective decision making: discrete collective estimation in large decision spaces. Swarm Intelligence **18**(2), 215–241 (2024)
49. Siemensma, T., Haghighat, B.: Optimization of collective Bayesian decision-making in a swarm of miniaturized vibration-sensing robots. Swarm Intelligence **20**, 1–32 (2026)
50. Soma, K., Vardharajan, V.S., Hamann, H., Beltrame, G.: Congestion and scalability in robot swarms: a study on collective decision making. In: International Symposium on Multi-Robot and Multi-Agent Systems (MRS). pp. 199–206. IEEE (2023)
51. Soorati, M.D., Krome, M., Mora-Mendoza, M., Ghofrani, J., Hamann, H.: Plasticity in collective decision-making for robots: Creating global reference frames,

detecting dynamic environments, and preventing lock-ins. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4100–4105. IEEE (2019)
52. Stillman, N.R., Kovacevic, M., Balaz, I., Hauert, S.: In silico modelling of cancer nanomedicine, across scales and transport barriers. npj Computational Materials **6**(92) (2020)
53. Strobel, V., Castelló Ferrer, E., Dorigo, M.: Managing Byzantine robots via blockchain technology in a swarm robotics collective decision making scenario. In: Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). pp. 541–549. IFAAMAS (2018)
54. Strobel, V., Castelló Ferrer, E., Dorigo, M.: Blockchain technology secures robot swarms: A comparison of consensus protocols and their resilience to Byzantine robots. Frontiers in Robotics and AI **7** (2020)
55. Strobel, V., Pacheco, A., Dorigo, M.: Robot swarms neutralize harmful Byzantine robots using a blockchain-based token economy. Science Robotics **8**(79), eabm4636 (2023)
56. Talamali, M.S., Saha, A., Marshall, J.A.R., Reina, A.: When less is more: Robot swarms adapt better to changes with constrained communication. Science Robotics **6**(56), eabf1416 (2021)
57. Valentini, G., Brambilla, D., Hamann, H., Dorigo, M.: Collective perception of environmental features in a robot swarm. In: Swarm Intelligence (ANTS). pp. 65–76. Springer, Cham (2016)
58. Valentini, G., Hamann, H., Dorigo, M.: Self-organized collective decision making: The weighted voter model. In: 13th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). pp. 45–52. IFAAMAS (2014)
59. Valentini, G., Hamann, H., Dorigo, M.: Efficient decision-making in a self-organizing robot swarm: On the speed versus accuracy trade-off. In: 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). pp. 1305–1314. IFAAMAS (2015)
60. Van Calck, L., Pacheco, A., Strobel, V., Dorigo, M., Reina, A.: A blockchain-based information market to incentivise cooperation in swarms of self-interested robots. Scientific Reports **13**, 20417 (2023)
61. Zakir, R., Carletti, T., Dorigo, M., Reina, A.: Bio-inspired decision making in swarms under biases from stubborn robots, corrupted communication, and independent discovery. arXiv:2509.07561 (2025)
62. Zakir, R., Dorigo, M., Reina, A.: Robot swarms break decision deadlocks in collective perception through cross-inhibition. In: International Conference on Swarm Intelligence (ANTS). LNCS, vol. 13491, pp. 209–221. Springer, Cham (2022)
63. Zakir, R., Dorigo, M., Reina, A.: Miscommunication between robots can improve group accuracy in best-of-n decision-making. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9014–9021. IEEE (2024)
64. Zakir, R., Salahshour, M., Dorigo, M., Reina, A.: Heterogeneity can enhance the adaptivity of robot swarms to dynamic environments. In: International Conference on Swarm Intelligence (ANTS), LNCS, vol. 14987, pp. 112–126. Springer, Cham (2024)