



Delft University of Technology

Shared Awareness Across Domain-Specific Artificial Intelligence An Alternative to Domain-General Intelligence and Artificial Consciousness

Deroy, Ophelia; Bacciu, Davide; Bahrami, Bahador; Della Santina, Cosimo; Hauert, Sabine

DOI

[10.1002/aisy.202300740](https://doi.org/10.1002/aisy.202300740)

Publication date

2024

Document Version

Final published version

Published in

Advanced Intelligent Systems

Citation (APA)

Deroy, O., Bacciu, D., Bahrami, B., Della Santina, C., & Hauert, S. (2024). Shared Awareness Across Domain-Specific Artificial Intelligence: An Alternative to Domain-General Intelligence and Artificial Consciousness. *Advanced Intelligent Systems*, 6(10), Article 2300740. <https://doi.org/10.1002/aisy.202300740>

Important note

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

Copyright

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

Takedown policy

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

Shared Awareness Across Domain-Specific Artificial Intelligence: An Alternative to Domain-General Intelligence and Artificial Consciousness

Ophelia Deroy, Davide Bacciu, Bahador Bahrami, Cosimo Della Santina, and Sabine Hauert**

Creating artificial general intelligence is the solution most often in the spotlight. It is also linked with the possibility—or fear—of machines gaining consciousness. Alternatively, developing domain-specific artificial intelligence is more reliable, energy-efficient, and ethically tractable, and raises mostly a problem of effective coordination between different systems and humans. Herein, it is argued that it will not require machines to be conscious and that simpler ways of sharing awareness are sufficient.

1. Introduction

There is a widespread belief that artificial intelligence (AI) should or will evolve to achieve general intelligence and even consciousness, somewhat mirroring an accelerated version of biological evolution. The feasibility and plausibility of this view are certainly debated^[1–5] but lacking are alternative visions where the progress of AI does not inherently require generality, nor eventually reaches human-like consciousness.

Here, we add our voice to a few ones (e.g., ref. [6]), which propose that investing in specialized AI systems tailored to specific tasks can be overall more effective than developing general intelligence. This competing vision is based on the diagnosis that,

de facto, many AI applications and users are currently working with fine-tuned or domain-specific AI, and that domain generality is not always required or reliable. It is also based on the idea that besides reliability, domain-specific systems can be more energy-efficient, and easier to comprehend and regulate.

Yet domain-specific systems also face a major challenge: how can systems designed to perform specialized roles, using heterogeneous languages and archi-

tectures, work together smoothly? The problem exists at all levels: in logistics, different specialized delivery systems operate with their own representation of space and goals, yet they need to coordinate their movement to work in the same space. On assembly lines, diverse robots or parts need to cooperate to sort, pack, and load fruits or products. In health care, domain-specific diagnostic AI systems can be optimized to analyze the patient's medical history and symptoms but also need to interact with other AI systems which assist in treatment recommendations or monitoring the patient's health.

Besides the problem of domain-specific AI systems collaborating with each other, another side is that human controllers or users need to interact not just with each domain-specific AI

O. Deroy
Faculty of Philosophy & Munich Center for Neuroscience
Ludwig Maximilian University
Geschwister-Scholl Platz 1, 80539 Munich, Germany
E-mail: ophelia.deroy@lrz.uni-muenchen.de

O. Deroy
Institute of Philosophy
School of Advanced Study
University of London
Senate House, London WC1E 7HU, UK

D. Bacciu
Computer Science Department
University of Pisa
L.go Bruno Pontecorvo 3, 56125 Pisa, Italy

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aisy.202300740>.

© 2024 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202300740

B. Bahrami
Faculty of Psychology
Ludwig Maximilian University
Gabelsbergerstrasse 62, 80333 Munich, Germany

C. Della Santina
Department of Cognitive Robotics
Delft University of Technology
Building 34, Mekelweg 2, 2628 CD Delft, Netherlands

C. Della Santina
Robotics and Mechatronics
German Aerospace Center (DLR)
82234 Wessling, Germany

S. Hauert
Bristol Robotics Laboratory
School of Engineering Mathematics and Technology
University of Bristol
BS81TW, UK
E-mail: sabine.hauert@bristol.ac.uk

but with the set or group and understand how they work together and can be aligned with their own human interests. In the examples above, a controller may want to check that all delivery AIs operate smoothly, rather than check each one and a patient may care about getting general health care, not separate domain-specific diagnoses and recommendations.

Our primary goal is concerned with lifting this overall collaboration challenge: can we ensure that the collaboration between domain-specific AI as well as humans in diverse roles works, as a plausible alternative to developing AI with general intelligence and eventually consciousness?

2. Two Evolutionary Paths: Domain-General Versus Domain-Specific Systems

The challenges bearing on AI can be likened to those faced by organisms stepping into new environments. Expanding into new domains means solving unprecedented challenges: what capabilities are needed to succeed in tasks like driving, motor control, cancer diagnosis, treatment recommendations, remote surgery, factory logistics, and nanorobotic drug delivery? What about generative tasks like vaccine discovery, musical creation, or robotic design?^[7]

Of course, AI and robotics operate under a different framework than living organisms, as they do not evolve through natural selection as described by Darwin. Instead, their progress is directed by human creators through deliberate planning and engineering. While natural evolution involves mutation, reproduction, and survival competition, AI and robotics evolve through hardware updates, computational advancements, economic demands, societal shifts, and ethical considerations.

Unlike living organisms, our choices have a direct and shorter-term influence on AI evolution. The expectations of researchers, regulators, and economic stakeholders, and our ways of envisioning or presenting solutions^[8] are driving what will come next.

Despite the difference between AI systems and living organisms many actors embrace the idea that AI systems need to go through a kind of accelerated version of biological evolution—ultimately developing general intelligence and eventually consciousness, which is variously seen as a byproduct or a precondition for general intelligence (see refs. [9,10] for overview). The idea here is not just to use computational models of consciousness to inspire solutions for AI—but to think that this transfer would make the system conscious.

Our core goal here is not to raise more objections to this goal than already exist, but to suggest that the evolutionary pressures at play in the expansion of AI systems do not create a demand for domain-general intelligence nor for the rich, integrated subjective experience that defines human consciousness. While many impressive current efforts are invested in the development of general intelligence, we have good reasons to place our efforts in domain-specific AI systems. Below we explain what these arguments are, in terms of reliability, lower energy costs, and ethical tractability, but also show what solutions lie ahead to enable the collaboration between heterogeneous systems, as well as with humans.

3. The Comparative Benefits of Domain-Specific AI

Tackling increasingly complex problems does not necessarily mean they are less domain-specified: many of the AI tools that are currently developed and adopted are meant to solve tasks within specific domains and tasks, be they expert-level ones (medical diagnosis), social ones (care robots, interactive conversations), or new ones operating at previously inaccessible scales (nanorobotics). There is however also a mounting pressure to link several domains (vision and image description; diagnosis and treatment).

Large language models (LLMs) are the closest so far to prefigure generalizable competencies across domains, and their occurrence seems to have shifted the hopes for AI closer in time.^[11] This said, they are still not wholly domain-general: to achieve good levels of performance in a task also requires fine-tuning to a domain and sometimes a different natural language. LLMs mostly solve tasks that use language and do less well in solving problems that require for instance social intelligence or complex spatial problems. Their capacity to perform new tasks is often dependent on extensive training and their capacity to cross-domains without training (known as zero-shot learning) remains challenging; achieving generalization to some new tasks can hinder their performance in others and the capacity to cross-domains remains dependent on real-world knowledge of specific domains.^[12]

How, when, and whether these problems can be fixed is a domain of speculation as much as proof. So far, the promises of domain-specific systems, including fine-tuned LLMs, are more concrete than those of machines capable of navigating across all domains and scales. Yet, with the hopes and promises of LLMs now wide open, continuing to bet on the superiority of domain-specific systems requires more arguments.

The first one is that domain-specific machines can be more effectively explained, controlled, and regulated. A domain-specific AI is considered more ethically tractable than a domain-general AI because it operates within well-defined tasks and domains, focusing on specific applications rather than attempting to emulate human intelligence across various areas. This specificity allows for easier oversight and control, reducing the risk of unintended consequences or ethical dilemmas that may arise from the opacity, complexity, and unpredictability of a domain-general AI system (e.g., ref. [13]; see also overview in ref. [14]). Governments and ethical boards are unable to regulate too broadly defined AI uses, but they can define where, when, and for whom particular AI can be ethically acceptable. Automated face recognition may be acceptable to assist custom officers in border control or help treat asylum requests more efficiently, but only within some boundaries and with appropriate high-security criteria. Endowing domain-general AI, also capable of speaking and taking health, educational, or military decisions with such capacities is a threat. Biases in training data and outputs are also more traceable and correctable for domain-specific AI. The field of Explainable AI also recognizes the need to deploy specific explanations for specific contexts and users.

Besides these ethical and social imperatives, another imperative bears also directly on AI and us alike: sustainable

development and environmental responsibility. Domain-specific AI promises greater sustainability and lower energy consumption. As a principle, operating a general-purpose system when a domain-specific system suffices is unlikely to be energy-efficient. The most general examples to date—the LLMs—demand substantial energy investments, not to mention extensive human oversight and safeguards. Evidence shows that multipurpose, generative architectures incur significantly higher energy and emission costs compared to task-specific systems across various tasks, even when controlling for the number of model parameters.^[15–17]

4. The Emerging Pressure to Collaborate

Expanding domain-specific AI raises new challenges. The multiplication of task- or domain-specific agents increasingly requires that different, heterogeneous agents not only coexist but also coordinate or even cooperate. Many concerns we commonly hear about—a car's inability to navigate a flooded road or a drone's hesitation between two equally perilous obstacles—are serious. But an equally serious one is the risk of automated vehicles or drones from different manufacturers colliding due to a lack of coordination in the spaces that they will de facto share. This collaboration problem is attracting increasing attention and is variously formulated as a problem of interoperability or that of multiagent cooperation.^[18]

Going back to the analogy with evolution, if AI evolution is developing more and more domain-specific systems, the pressure is not only to master and adapt to new environments, but share the same ones with other, sometimes very different “species” of AI, as well as with us, humans. Without a way to solve this collaboration challenge, a clear risk is that of arrested development, with each AI bound to its own local niche. A second risk is that of eventual competition or chaos as different brands and systems do not manage to operate together. The energy savings realized by keeping tasks specific at the unit level may be real, but they need to demonstrate their capacity to scale up when it comes to solving complex coordinated tasks.

The challenge also poses an additional problem: can humans comprehend or monitor not just the functioning of individual systems but also their interactions?^[19] Human acceptance is not just one constraint among many to be optimized, alongside factors like effectiveness, energy efficiency, speed, and robustness; it is a prerequisite for any system to be produced and disseminated, where simple solutions too quickly based on human cooperation can eventually backfire.^[20]

5. How to Solve the Collaboration Problem?

Does the challenge of collaboration not lead us also to reach a similar point where machine consciousness will be key to AI evolution? Would it not also be necessary for domain-specific AI systems to be conscious to be able to flexibly coordinate with each other, and be sufficiently transparent to humans?

Current theories about the functions of consciousness do not see it that way. Despite differences between theories and perspectives, there is a widespread consensus—going back at least to Sherrington^[21]—that the key function of consciousness comes

with the integration of information across various systems but still strictly within an organism. This “integration consensus”^[22] at the level of function is visible in Baars’ “global workspace” theory,^[23,24] Dehaene’s Global Neuronal Workspace^[25] as well as Edelman (e.g., refs. [26–28]). The place and way this integration works are subject to ongoing debates, yet most proposals suggest that consciousness is necessary for making information available across various systems to guide the selection of a single goal or course of action for the organism.^[2,29,30] Relatedly, most theories see biological consciousness as an evolved individual capacity, whose advantages or role lies primarily in optimizing the behavior of single individuals. This does not mean that consciousness does not have benefits further down the line for communication and social coordination but that these benefits are generally not seen as the problem it evolved to solve (see, refs. [31–33] for discussion). Philosophically, consciousness is also primarily viewed as an individual phenomenon. Features like privacy and self-reference are recognized as integral to consciousness.^[34]

When the challenge at hand involves orchestrating numerous simultaneous or sequential actions across different AI-specialized systems, the presence of consciousness as a private integrative mechanism within each system is neither essential nor sufficient. What is needed is the capacity for selectively sharing relevant states with other AI systems to facilitate coordination and cooperation—or collaborative shared awareness for short. As the word “awareness” is sometimes used as a synonym for consciousness, it is important to see why collaborative awareness is significantly different from consciousness. While there may be other differences, we want to stress three main ones.

First, shared awareness is not a private state, by definition. If a swarm of bots has a shared awareness of the whole factory floor, this shared awareness is not reducible to the representation of space that each individual agent has. It is an emergent property. A state of consciousness is private to each agent in two senses: the experience is uniquely enjoyed by the subject, from her perspective; the subject has immediate and privileged access to her conscious experience. Collaborative awareness is radically different from this “private theatre” of consciousness because it is held between agents, and offers the same access to all.

Second, shared awareness can be only transient, while consciousness is continuous. Once an organism is capable of consciousness, its consciousness depends on its state of wakefulness as well as on the type of inputs: consciousness is not permanent but it is a continuous state. Collaborative awareness only selectively shares states with others when there is a need to coordinate individual goals or cooperate on a common goal—for instance, sharing will occur when another AI system comes in close proximity, or when a certain goal is selected.

Third, shared awareness can be selective. In a swarm of bots again, each agent may be able to represent both its energy levels, its goals, and the space around it. For shared awareness, only space may be relevant. While the dominant views of consciousness mean it is integrated or unified, shared awareness can be partitioned across different agents: one system may need to share spatial information with another system, energy levels with their controller, and other aspects such as their confidence with other systems or their users.

The ability to share and report what one is aware of seeing, thinking or planning is often associated with conscious states,

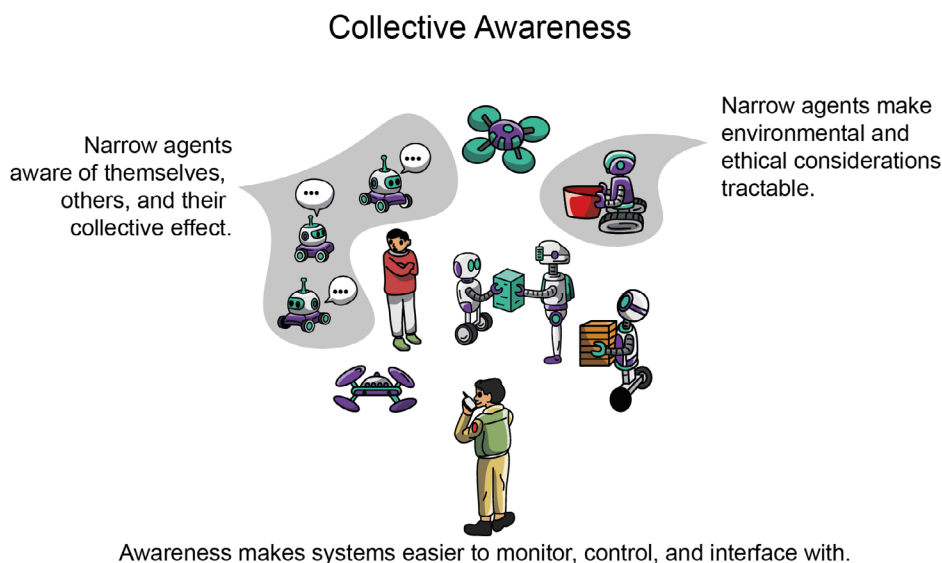


Figure 1. Features of collective awareness.

but by no means necessary for consciousness: patients with lock-in syndrome or minimal states of consciousness can lose the capacity to report, but still be conscious; nonhuman animals can be conscious without being able to report their states; in experiments, humans also can have a richer array of states than the ones they can report.^[35,36]

6. Future-Looking Perspectives

Various researchers have expressed hopes, fears, optimism, or skepticism regarding artificial consciousness. Opinion leaders and the media have amplified these discussions, viewing “machine consciousness” as fraught with significant ethical and anthropocentric dilemmas. Many other AI researchers choose to steer clear of the consciousness debates, considering it a too intricate and opaque concept. Debates and speculation are of course integral to scientific inquiry and essential in democratic societies.

Our point is that debating consciousness for machines already accepts a certain view of the evolution of AI. Whether human or biological domain-general intelligence goes hand in hand with a form of integration of information that requires simulating or embedding consciousness in agents is relatively accepted. Yet achieving domain-general intelligence is not the only path for AI and collaboration across domain-specific systems and with humans can be as efficient. It is also perhaps cheaper and more ethical.

Robot swarms are an example of a cheap, flexible system where many actors can work together to deliver a service. Yet individual robots are typically only aware of their local environment, nearby robots in the swarm, and the humans they interact with.^[37] This makes it challenging for robots to know if they are successfully contributing to the overall system, to communicate the swarm state to human users and operators, and to enable humans to interact with the swarm. Engineering collaborative awareness into the swarm could make them easier to deploy,

monitor, and control. This could be done by enabling sharing and consolidation of information related to the services they are meant to accomplish, and unifying how such information is presented and interfaced with toward making them trustworthy and actionable.^[38]

The framework of collaborative awareness means that such collective awareness does not stop at the local scale of a given swarm or for a set of robotic agents engineered by a single brand or serving a given mission, but can also emerge when coordination or cooperation across systems is needed (**Figure 1**). To resort to an analogy, systems should have a way to share their use-relevant states or plans with other systems as a kind of “selective telepathy.” The fact that consciousness in humans does not allow private states to be shared should not limit our engineering ideas, and thinking of awareness as a state that is fundamentally shared across agents—and not present in individuals but only selectively emerging in group interactions—represents a framework to enable heterogeneous agents to collaborate.

Acknowledgements

The authors acknowledge support from the European Union under grant agreement 101070918, EMERGE project. UK participants in Horizon Europe Project 101070918 are supported by UKRI grant no. 10038942. S.H. is supported by the ESRC Centre for Sociodigital Futures grant no. ES/W002639/1.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

collective awareness, human–AI Interactions, machine consciousness, sustainability, swarm robotics

Received: November 8, 2023
Revised: May 14, 2024
Published online:

- [1] J. Aru, M. E. Larkum, J. M. Shine, *Trends Neurosci.* **2023**, *46*, 1008.
- [2] S. Dehaene, H. Lau, S. Kouider, in J. S. von Braun, M. Archer, G. M. Reichberg, M. Sánchez Sorondo *Robotics, AI, and Humanity: Science, Ethics, and Policy*, Springer, Cham. **2021** pp. 43–56.
- [3] J. Kleiner, T. Ludwig, If Consciousness is Dynamically Relevant, Artificial Intelligence Isn't Conscious, arXiv:2304.05077, **2023**.
- [4] E. Hildt, *AJOB Neurosci.* **2023**, *14*, 58.
- [5] T. Metzinger, *J. Artif. Intell. Conscious.* **2021**, *8*, 43.
- [6] S. Pal, M. Bhattacharya, S. S. Lee, C. Chakraborty, *Ann. Biomed. Eng.* **2024**, *52*, 451.
- [7] F. Stella, C. Della Santina, J. Hughes, *Nat. Mach. Intell.* **2023**, *5*, 561.
- [8] O. Deroy, *Topoi* **2023**, *42*, 881–889.
- [9] R. Fjelland, *Humanit. Soc. Sci. Commun.* **2020**, *7*, 1.
- [10] A. Juliani, K. Arulkumaran, S. Sasai, R. Kanai, On the Link Between Conscious Function and General Intelligence in Humans and Machines, arXiv preprint arXiv:2204.05133, **2022**.
- [11] K. Grace, H. Stewart, J. F. Sandkühler, S. Thomas, B. Weinstein-Raun, J. Brauner, Thousands of AI authors on the future of AI, arXiv:2401.02843, **2024**.
- [12] R. Kirk, A. Zhang, E. Grefenstette, T. Rocktäschel, *J. Artif. Intell. Res.* **2023**, *76*, 201.
- [13] J. Stenseke, *Artif. Intell. Rev.* **2024**, *57*, 1.
- [14] J. J. Bryson, in *The Oxford Handbook of Ethics of AI*, Oxford University Press, New York, NY **2020** pp. 1–25.
- [15] C. J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, C. Gloria, F. A. Behram, J. Huang, C. Bai, M. Gschwind, A. Gupta, M. Ott, A. Melnikov, S. Candido, D. Brooks, G. Chauhan, B. Lee, H.-H. S. Lee, B. Akyildiz, M. Balandat, J. Spisak, R. Jain, M. Rabbat, K. Hazelwood, *Proc. Mach. Learn. Syst.* **2022**, *4*, 795.
- [16] K. Chadli, G. Botterweck, T. Saber, in *Proc. of the 4th Workshop on Machine Learning and Systems*, Athens, Greece **2024** pp. 200–207.
- [17] A. S. Luccioni, Y. Jernite, E. Strubell, Power Hungry Processing: Watts Driving the Cost of AI Deployment? arXiv:2311.16863, **2023**.
- [18] M. Noura, M. Atiquzzaman, M. Gaedke, *Mobile Networks Appl.* **2019**, *24*, 796.
- [19] D. Bacciu, S. Akarmazyan, E. Armengaud, M. Bacco, G. Bravos, C. Calandra, E. Carlini, A. Carta, P. Cassara, M. Coppola, C. Davalas, P. Dazzi, M. C. Degennaro, D. D. Sarli, J. Dobaj, C. Gallicchio, S. Girbal, A. Gotta, R. Groppox, V. Lomonaco, G. Macher, D. Mazzei, G. Mencagli, D. Michail, A. Micheli, R. Peroglio, S. Petroni, R. Potenza, F. Pourdanesh, C. Sardanios, et al., in *IEEE Inter. Conf. on Omni-Layer Intelligent Systems (COINS)*, IEEE, Piscataway, NJ **2021**, pp. 1–6.
- [20] J. Karpus, A. Krüger, J. T. Verba, B. Bahrami, O. Deroy, *Iscience* **2021**, *24*, 102679.
- [21] C. S. Sherrington, in *The Integrative Action of The Nervous System*, Yale University Press, New Haven, CT **1906**.
- [22] E. Morsella, *Psychol. Rev.* **2005**, *112*, 1000.
- [23] B. J. Baars, in *A Cognitive Theory of Consciousness*, Cambridge University Press, New York, NY **1988**.
- [24] B. J. Baars, *Trends Cogn. Sci.* **2002**, *6*, 47.
- [25] S. Dehaene, J. P. Changeux, L. Naccache, in *Characterizing Consciousness: From Cognition to the Clinic?* **2011**, pp. 55–84.
- [26] G. M. Edelman, in *The Remembered Present*, Basic Books, New York, NY **1989**.
- [27] G. M. Edelman, *Proc. Natl. Acad. Sci. U S A* **2003**, *100*, 5520.
- [28] G. Tononi, *BMC Neurosci.* **2004**, *5*, 42.
- [29] N. Block, *Behav. Brain Sci.* **1995**, *18*, 227.
- [30] O. Deroy, N. Faivre, C. Lunghi, C. Spence, M. Aller, U. Noppeney, *Multisens. Res.* **2016**, *29*, 585.
- [31] C. D. Frith, in *Frontiers Of Consciousness: Chichele Lectures* (Eds: L. Weiskrantz, M. Davies), Oxford University Press, Oxford, England **2008**, pp. 225–244.
- [32] C. D. Frith, *Cognit. Neurosci.* **2011**, *2*, 117.
- [33] N. Shea, A. Boldt, D. Bang, N. Yeung, C. Heyes, C. D. Frith, *Trends Cognit. Sci.* **2014**, *18*, 186.
- [34] C. W. Tyler, *Front. Psychol.* **2020**, *11*, 521207.
- [35] N. Block, *Trends Cognit. Sci.* **2011**, *15*, 567.
- [36] D. Derrien, C. Garric, C. Sergeant, S. Chokron, *Neurosci. Conscious.* **2022**, 2022, niab043.
- [37] S. Jones, E. Milner, M. Sooriyabandara, S. Hauert, *Adv. Intell. Syst.* **2020**, *2*, 2000110.
- [38] J. Wilson, G. Chance, P. Winter, S. Lee, E. Milner, D. Abeywickrama, S. Windsor, J. Downer, K. Eder, J. Ives, S. Hauert, in *Proc. of the First International Symp. on Trustworthy Autonomous Systems*, Edinburgh, UK **2023** pp. 1–11.