Visual Model Building for Robot Sensemaking: Perspectives, Challenges, and Opportunities

Agnese Chiatti[1], **Gianluca Bardaro**[2], **Matteo Matteucci**[2], **Enrico Motta**[1]

[1] Knowledge Media Institute, The Open University (UK)
[2] AI & Robotics Lab, Politecnico di Milano (Italy)
agnese.chiatti@open.ac.uk, gianluca.bardaro@polimi.it, matteo.matteucci@polimi.it, enrico.motta@open.ac.uk

Abstract

Robots can help with many visually-intense and onerous tasks that are traditionally carried out by human workers, such as the inspection of critical infrastructures or the management of crops. However, before we can safely delegate tasks to robots, we ought to ensure that they can reliably make sense of the environments in which they are deployed. To this aim, building world models that resemble the complexity of the real world is critical. Despite research efforts in the AI and Robotics communities towards tackling the problem of model building, few works exist that approach this problem by considering perspectives and lessons learned from both fields. In this position paper, we use three strategic application domains in Robotics to argue for the centrality of visual model building to support robot sensemaking.

Introduction and Background

Robots can help with many daily activities by taking over inconvenient or unsafe tasks. Indeed, they are already being deployed across many economic sectors, including warehouse automation and last-mile deliveries (Chen et al. 2021), waste disposal (Samonte, Baloloy, and Datinguinoo 2021), and assistance to elderly people (Bardaro, Antonini, and Motta 2022), to name just a few. In assistive scenarios, robots can help preventing risks to the Health and Safety of fragile patients and older adults, such as tripping hazards, or fire threats - e.g., a pot left on the stove. In agricultural settings, robots can be deployed for the autonomous management of crops (Bertoglio et al. 2021a). For example, they can help estimating more precisely the amount of water and herbicides needed for farming, thus mitigating both resource waste and environmental pollution[1].

To operate successfully in complex and dynamic environments, service robots are expected to reliably interpret the data collected through their perceptual sensors (Alatise and Hancke 2020). This process of understanding and reconciliation of different experiences is also known as *sensemaking* (Russell et al. 1993).

Back to the scenario of a pot left on the stove, a robot would need to first detect the pot and the stove. It would also need to recognise that the two objects are in contact with one another. Importantly, it would need to know that pots are food containers and that food can burn if heated for a prolonged period of time. Similarly, a robot that monitors crops needs access to a wide range of capabilities. It ought to be capable of recognising different plant species. It also has to know which actions are most appropriate in the different phases of the plant lifecycle. Crucially, it ought to deal with environments that are characterised by a high variety of seasonal effects and weather changes.

As conveyed by these examples, several different types of knowledge and reasoning capabilities are required for robots to make sense of complex and dynamic environments. These include, but are not limited to spatial, causal, and temporal reasoning skills. Moreover, when robots are deployed in authentic problem-solving scenarios, they must be able to learn from their experiences, to asses the typicality and plausibility of a situation. That is, they must exhibit some degree of commonsense (Davis and Marcus 2015; Levesque 2017).

From the standpoint of visual perception, robots ought to use not only their vision system, but also their reasoning system, and external knowledge sources to make sense of the environment. In our previous work (Chiatti, Motta, and Daga 2020), we used the term *visual intelligence* to describe this process of visual sensemaking and identified the epistemic requirements that contribute to a robot's visual intelligence. A key epistemic component that has emerged as transversal to all the requirements in (Chiatti, Motta, and Daga 2020) is the capability of *model building*. Namely, to achieve visual sensemaking in real-world scenarios, robots need to maintain a fine-grained cognitive model of their surroundings (Lake et al. 2017). In Robotics, the need for persistent models that can bridge the gap between sub-symbolic perceptual data and symbolic representations has motivated the proposal of representational models known as *semantic maps*. Semantic maps contain "in addition to spatial information about the environment, assignments of mapped features to entities of known classes" (Nüchter and Hertzberg 2008). Thus, the semantic mapping task is closely intertwined with the issue of grounding successive observations of the same object entity, also known as the *object anchoring* problem (Coradeschi and Saffiotti 2003; Bonarini, Matteucci, and Restelli 2001).

On the one hand, Deep Learning (DL) methods have ex-

[1] as prescribed by the EU Farm to Fork strategy: https://food.ec.europa.eu/horizontal-topics/farm-fork-strategy_en

pedited the symbol grounding problem, by automating the discovery of features and patterns from unstructured perceptual data. On the other hand, these methods are unfit for model building (Lake et al. 2017). Indeed, DL models are generally incapable of retaining the information learned previously - i.e., *catastrophic forgetting* (Lesort et al. 2020). Their ability to learn from scarce and sparsely distributed data and to generalise beyond the data distribution learned during training is also inherently limited (Goyal and Bengio 2022; Lesort et al. 2020). Moreover, representations learned through DL are difficult to discern, decompose, and reuse across tasks - an issue also known as *disentanglement* (Hu et al. 2021).

By contrast, sensemaking methods that are based on reasoning over symbolic knowledge representations provide more explainable and reusable components for modelling the rules (physical, causal, procedural, etc.) that govern the robot's environment and for maintaining a persistent record of the robot's experiences. However, purely symbolic methods rely on the availability of explicit knowledge representations to support the reasoning process. As analysed in (Chiatti, Motta, and Daga 2020), state-of-the-art, large-scale Knowledge Bases only provide a partial coverage of the knowledge sources (e.g., physical, causal, spatial, commonsense-based) that are needed by an embodied agent for visual sensemaking.

Thus, from the perspective of model building, DL methods and symbolic methods exhibit complementary strengths. Under similar premises, a recent trend in AI has been to integrate symbolic and sub-symbolic models for producing *hybrid* (Aditya, Yang, and Baral 2019) or *neuro-symbolic* learning methods (Sarker et al. 2022). In this view, semantic maps provide a prime example of neuro-symbolic model, which synthesises the inputs of different DL-based and knowledge-based components. Despite this evidence, the neuro-symbolic learning and semantic mapping fields have evolved almost independently of one another. In this position paper, we argue that a systematic framework for visual model building is needed, which integrates contributions and lessons learned from both fields. To illustrate the benefits of introducing the envisaged framework to support works at the intersection of AI and Robotics, we will examine three application domains that have been recently identified by the European Commission as particularly strategic.

Application Domains

To illustrate the potential impact of the requirement of visual model building, we consider three of the four application domains that the European Commission has highlighted as prioritary in the programme for Research and Innovation Actions in Robotics[2] - last updated in April 2022. First, we discuss the case of Health&Safety monitoring, which falls under the Priority Area (PA) of inspection and maintenance of infrastructure. Moreover, we examine the case of assisting patients and older adults at home, which pertains the PA of healthcare. Lastly, we discuss the domain of precision farming, which concerns the agri-food PA. We exclude from our analysis the PA of agile production to focus on environments that are more unconstrained than manufacturing scenarios.

[2]https://cordis.europa.eu/programme/id/H2020_ICT-46-2020

Autonomous Inspection and Maintenance

A first compelling use-case for the deployment of robotic assistants is that of monitoring the Health and Safety (H&S) of critical infrastructures. To this aim, at the Knowledge Media Institute (KMi), we have developed a prototype Health and Safety robot inspector, HanS. HanS is expected to check that office spaces comply with a set of Health and Safety rules. For instance, it ought to recognise tripping hazards, such as cables dangling in the middle of corridors, and it must also carry out regular checks to verify, for example, that fire extinguishers are correctly mounted at their dedicated locations. HanS uses Deep Learning methods for recognising objects, but is also equipped with reasoning modules that consider the typical size and spatial relations of objects to refine the DL predictions (Chiatti, Motta, and Daga 2022). Importantly, HanS can keep track of its observations over time through a semantic map, where predictions recorded at the same location are anchored to the same object instance. These grounded predictions are then used for generating a graph that models the spatial relations between neighbour objects. The graph is further enriched with commonsense facts extracted automatically from the Quasimodo Knowledge Base (Romero and Razniewski 2020). Having access to a knowledge-enriched world model, HanS can identify the fire hazard shown in Figure 1, where a heater lies next to a pile of books. Namely, it can infer that books are made of paper, that paper is a flammable material, and that electric heaters can cause a fire (Figure 1b).

The HanS prototype demonstrates that equipping robots with methods for anchoring successive observations and linking perceptual data with semantic and symbolic knowledge is essential to enable higher-level reasoning and sensemaking. Nonetheless, the model building system embedded in the HanS application could be extensively improved. First, handling objects that are removed or disappear from view (e.g., because they are occluded by other objects) is a long-standing problem in semantic mapping, which has motivated the proposal of probabilistic approaches to model uncertain object positioning (Persson et al. 2020). Moreover, multiple object attributes could be considered for anchoring purposes, in addition to the object location - e.g., the object colour and classification predictions, as in (Günther et al. 2018). One unexplored component that could significantly contribute to refining anchoring systems is considering the spatio-temporal evolution of object attributes for making anchoring decisions. For example, static objects do not need to be re-recorded and re-identified as frequently as moving objects, thus making the anchoring process more efficient.

Another crucial direction of improvement, in this context, is evaluating the ability of inspection robot developed indoors to support the exploration of sensory-deprived environments. Underground maintenance tunnels, which are characterised by low lighting conditions and by the lack of GPS signal, are one example of sensory-deprived setting. In such contexts, the availability of reliable world models is particularly important, because it can help compensating

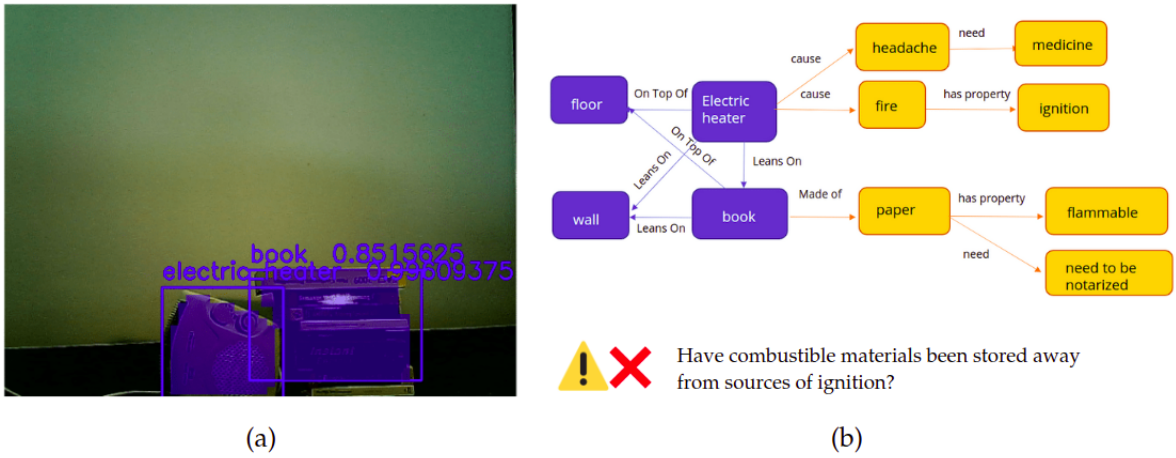

Figure 1: (a) An electric heater is spotted near a pile of books. (b) Because a flammable object is in contact with an ignition source, this scenario represents a potential fire hazard.

for the robot's perceptual deprivation. For instance, the annotation of semantic landmarks, such as hatches, pipelines, and ladders, can augment the data collected through camera and laser sensors, to support Simultaneous Localisation and Mapping (SLAM) and obstacle avoidance tasks (Hong et al. 2021).

Home Healthcare

The assistance of older adults at home is a very actively-researched area in Service Robotics for healthcare. Indeed, mild cognitive (or visual) impairments are leading causes of incidents in patients' homes (Bardaro, Antonini, and Motta 2022). For instance, elderly patients may trip over small objects left on the floor and fall, or leave the stove on due to forgetfulness. They may also need help retrieving personal belongings that have been misplaced, such as their glasses, wallet, or house keys. In such scenarios, a persistent recording of previous observations gives the robot access to the last known location of personal items.

Similarly to the HanS' use-case, the assistive scenario implies that the robot perceptions are linked to a diversified set of knowledge sources and symbolic reasoning modules. For example, to recognise that a pot has been forgotten on the stove, the robot would need causal and spatio-temporal reasoning capabilities - e.g., how long has the stove been on for? Moreover, it needs to be capable of discriminating between harmless activities (the stove is on only as long as it is needed for cooking) and hazardous, anomalous events (the stove has been left on for a long time). In this scenario, the robot's sensemaking is highly dependendent on the patient's routines and preferences compared to the HanS' case, where the set of rules to be checked is known a priori. Here, the interaction between the robot, the user, and the environment is a crucial contributing factor to the model building process.

In the context of the EU-funded project GATEKEEPER, we are exploring assistive scenarios where the robot is embedded in a wider system of Smart Home sensors (Bardaro, Antonini, and Motta 2022). In this configuration, a unified model of the world can help tracking and reconciling the data that is gathered through multiple input sources.

Precision farming

The deployment of robots in agricultural settings can improve the sustainability of traditional practices for crop management, to account for climate change, by reducing the amount of water, herbicides, and pesticides that are used in farming, as well as the emission of greenhouse gas (Bertoglio et al. 2021a).

For tending to crops, a robot needs to keep track of the characteristics and state of plants, i.e., a capability also known as plant phenotyping. It also needs access to expert knowledge about which actions are appropriate in the different phases of the plant lifecycle. However, crops are particularly challenging operational environments for robots, because they are characterised by significant weather and seasonal variations.

This challenging scenario calls for methods for building knowledge representations that can model uncertainty and that can be dynamically updated to reflect environmental changes. Approaches like the one presented in (Deeken, Wiemann, and Hertzberg 2018) use semantic maps for annotating the different field boundaries to orchestrate fleets of roboticised vehicles. However, the application of symbolic representations to augment sensor processing in challenging outdoor environments is still relatively unexplored.

In this context, a longitudinal recording of agricultural environments would provide the opportunity to extend state-of-the-art methodologies with different respects. Indeed, conducting a longitudinal study of environments characterised by rapid change offers the opportunity to test the generalisation ability of models developed in scenarios that are relatively more static and constrained. Specifically, precision farming solutions require a timely update not only of the robot's model of the crop, but also of the robot's learning models. As such, tackling the problem of incremental model building in this domain contributes to extending state-of-the-

art methods for Domain Adaptation and Continual Learning (Lesort et al. 2020; Lee et al. 2022; Evci et al. 2022).

Conclusion and Open Challenges

The capability of building and maintaining world models that resemble the complexity of real-world environments is essential for robots to operate effectively in a variety of application scenarios. As emerged through the analysis of the three application domains presented in this paper, defining a systematic framework for model building raises several different challenges.

Knowledge Representations for Robots. The large-scale Knowledge Bases that are currently available are vastly limited in their support of robotic tasks. The commonsense knowledge required for robots to reason on the plausibility of situations (Levesque 2017) is particularly difficult to formalise and explicitly encode, because it has never been explicitly taught to us (Davis and Marcus 2015). Specifically, the analysis presented in (Chiatti, Motta, and Daga 2020) indicated that state-of-the-art comprehensive Knowledge Bases only provide a partial modelling of the intuitive physical properties of objects (e.g., their size, natural orientation and symmetry), their composing parts, as well as their typical locations and everyday uses. In some cases, the required properties are only provided for a limited set of objects. In other cases, the taxonomies used for modelling the knowledge attributes are inconsistent or completely unstructured. Particularly striking is the lack of representations of the typical motion trajectories of objects (as static, moving, or movable), which would significantly aid model building tasks. These gaps call for the design of novel methods for providing the knowledge sources required for robot sensemaking by either re-purposing the existing sources, engineering the missing knowledge *ex novo*, or autonomously constructing the missing knowledge from the bottom up, by learning through repeated observation.

Handling change and unpredictability. Building realistic world models requires to ground semantic concepts to high-volume perceptual observations. Handling objects that may be removed or disappear from view is an open problem in semantic mapping (Persson et al. 2020; Han et al. 2021). Moreover, this problem domain calls for robust methods for modelling change and for periodically checking the temporal-validity of the acquired information. One fundamental problem of modelling real-world scenarios is that the knowledge available about the environment (e.g., about the classes of objects and events that compose the environment) is incomplete and cannot be modelled exclusively a priori, in a declarative fashion. As such, the sole improvement of existing knowledge sources is not sufficient for robots to operate reliably in open-ended, complex environments. If we aim at deploying robots in real-world environments, we also ought to rethink our experimental design and benchmarking practices. The predominant approach in the DL community is to evaluate a system's performance on benchmark datasets and on simulated environments, thus broadening the disconnect between AI contributions and Robotics practices. As an alternative to controlled experiments in constrained and/or simulated scenarios, Schiaffonati (2022) proposes *explorative experiments*, where robots are evaluated iteratively on the basis of their repeated interaction with real-world environments, to refine their fulfilment of a specific task. A promising research direction, in this context, is the development of reproducible benchmarking frameworks for evaluating robots through competitions (Bertoglio et al. 2021b).

Multi-modal knowledge acquisition. In embodied sensemaking systems, knowledge is acquired in at least two directions: (i) background knowledge retrieved externally informs the robot's interpretation of the environment, and (ii) the observation of the environment also helps refining the robot's sensemaking. This scenario is further complicated in cases where additional environmental sensors complement the robot's observations, and in the case of multi-agent systems. As such, methods for visual model building ought to address the problem of integrating the information collected through different sensors and embodied agents, by providing modular architectural components that rely on interoperable, unified data representations. Ultimately, the humans who collaborate and interact with the robot are another key source of information about the environment (Han et al. 2021). Therefore, the task of visual model building also requires an understanding of data modalities that span beyond visual perception. Robust speech recognition, gesture recognition and Natural Language Processing (NLP) are key capabilities in this context.

Meta-reasoning and meta-learning. Ultimately, because different reasoning components (e.g., physical, causal, spatial, temporal) and data modalities contribute to a robot's model building, robots are expected to autonomously reconcile different reasoning processes (meta-reasoning), and are also expected to learn how to learn (meta-learning). in settings characterised by frequent visual domain shifts, as in the agricultural domain, improved and adaptive learning methods are required. Recent work has indicated that selecting features and representations from different layers of DL models can improve the robustness to visual domain shifts (Lee et al. 2022; Evci et al. 2022), compared to the common practice of fine-tuning only the last layer of a DL model for transferring the knowledge learned on large-scale datasets on a few newly-acquired examples. However, these findings are yet to be validated on datasets collected through robotic sensors in real-world settings. More broadly, to interpret complex events, a robot ought to leverage the outputs of multiple reasoners, which may generate conflicting outcomes, especially in environments characterised by severe unpredictability and sensor deprivation.

Future work targeted at these challenging problems opens up significant opportunities of collaboration and knowledge exchange at the intersection of AI and Robotics.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation program through grant agreement No 857223 and from the Italian L'Oreal-UNESCO program "For Women in Science".

References

Aditya, S.; Yang, Y.; and Baral, C. 2019. Integrating knowledge and reasoning in image understanding. In *Proceedings of the Joint Conference on Artificial Intelligence (IJCAI)*, 6252–6259. IJCAI.

Alatise, M. B.; and Hancke, G. P. 2020. A review on challenges of autonomous mobile robot and sensor fusion methods. *IEEE Access*, 8: 39830–39846.

Bardaro, G.; Antonini, A.; and Motta, E. 2022. Robots for elderly care in the home: A landscape analysis and co-design toolkit. *International Journal of Social Robotics*, 14(3): 657–681.

Bertoglio, R.; Corbo, C.; Renga, F. M.; and Matteucci, M. 2021a. The digital agricultural revolution: A bibliometric analysis literature review. *IEEE Access*, 9: 134762–134782.

Bertoglio, R.; Fontana, G.; Matteucci, M.; Facchinetti, D.; Berducat, M.; and Boffety, D. 2021b. On the Design of the Agri-Food Competition for Robot Evaluation (ACRE). In *2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 161–166. IEEE.

Bonarini, A.; Matteucci, M.; and Restelli, M. 2001. Anchoring: do we need new solutions to an old problem or do we have old solutions for a new problem. In *Proceedings of the AAAI Symposium on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, 79–86.

Chen, C.; Demir, E.; Huang, Y.; and Qiu, R. 2021. The adoption of self-driving delivery robots in last mile logistics. *Transportation Research Part E: Logistics and Transportation Review*, 146: 1–16.

Chiatti, A.; Motta, E.; and Daga, E. 2020. Towards a framework for Visual Intelligence in Service Robotics: epistemic requirements and gap analysis. In *Proceedings of the International Conference on Principles of Knowldge Representation and Reasoning (KR)*, 905–916. IJCAI.

Chiatti, A.; Motta, E.; and Daga, E. 2022. Robots with Commonsense: Improving Object Recognition through Size and Spatial Awareness. In *Proceedings of the AAAI Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE)*. CEUR.

Coradeschi, S.; and Saffiotti, A. 2003. An introduction to the anchoring problem. *Robotics and autonomous systems*, 43(2-3): 85–96.

Davis, E.; and Marcus, G. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9): 92–103.

Deeken, H.; Wiemann, T.; and Hertzberg, J. 2018. A spatio-semantic model for agricultural environments and machines. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 589–600. Springer.

Evci, U.; Dumoulin, V.; Larochelle, H.; and Mozer, M. C. 2022. Head2Toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, 6009–6033. PMLR.

Goyal, A.; and Bengio, Y. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society*, 478(2266): 20210068.

Günther, M.; Ruiz-Sarmiento, J. R.; Galindo, C.; González-Jiménez, J.; and Hertzberg, J. 2018. Context-aware 3D object anchoring for mobile robots. *Robotics and Autonomous Systems*, 110: 12–32.

Han, X.; Li, S.; Wang, X.; and Zhou, W. 2021. Semantic mapping for mobile robots in indoor scenes: a survey. *Information*, 12(2): 92.

Hong, J.; de Langis, K.; Wyethv, C.; Walaszek, C.; and Sattar, J. 2021. Semantically-aware strategies for stereo-visual robotic obstacle avoidance. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2450–2456. IEEE.

Hu, J.; Cao, L.; Tong, T.; Ye, Q.; Zhang, S.; Li, K.; Huang, F.; Shao, L.; and Ji, R. 2021. Architecture disentanglement for deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 672–681.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Lee, Y.; Chen, A. S.; Tajwar, F.; Kumar, A.; Yao, H.; Liang, P.; and Finn, C. 2022. Surgical Fine-Tuning Improves Adaptation to Distribution Shifts. *arXiv preprint arXiv:2210.11466*.

Lesort, T.; Lomonaco, V.; Stoian, A.; Maltoni, D.; Filliat, D.; and Díaz-Rodríguez, N. 2020. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58: 52–68.

Levesque, H. 2017. *Common Sense, the Turing Test, and the Quest for Real AI*. The MIT Press.

Nüchter, A.; and Hertzberg, J. 2008. Towards semantic maps for mobile robots. *Robotics and Autonomous Systems*, 56(11): 915–926.

Persson, A.; Zuidberg Dos Martires, P. M.; De Raedt, L.; and Loutfi, A. 2020. ProbAnch: a modular probabilistic anchoring framework. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) - Demo track*, 5285–5287. IJCAI.

Romero, J.; and Razniewski, S. 2020. Inside quasimodo: Exploring construction and usage of commonsense knowledge. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 3445–3448. ACM.

Russell, D. M.; Stefik, M. J.; Pirolli, P.; and Card, S. K. 1993. The Cost Structure of Sensemaking. In *Proceedings of the ACM INTERACT and CHI Conference on Human Factors in Computing Systems*, 269–276. ACM.

Samonte, M. J. C.; Baloloy, S. H.; and Datinguinoo, C. K. J. 2021. e-TapOn: Solar-Powered Smart Bin with Path-based Robotic Garbage Collector. In *Proceedings of the IEEE International Conference on Industrial Engineering and Applications (ICIEA)*, 181–185. IEEE.

Sarker, M. K.; Zhou, L.; Eberhart, A.; and Hitzler, P. 2022. Neuro-symbolic artificial intelligence: Current Trends. *AI Communications*, 34: 197–209.

Schiaffonati, V. 2022. Explorative Experiments: A Paradigm Shift to Deal with Severe Uncertainty in Autonomous Robotics. *Perspectives on Science*, 30(2): 284–304.