# Machine Olfaction in Artificial Intelligence and Robotics

Kordel K. France
The University of Texas at Dallas
Richardson, TX, USA
kordel.france@utdallas.edu

Ovidiu Daescu
The University of Texas at Dallas
Richardson, TX, USA
ovidiu.daescu@utdallas.edu

## Abstract

Machine olfaction—the artificial replication of the sense of smell—faces significant challenges due to the absence of large, standardized training datasets. Unlike vision, language, and audio models, which benefit from extensive corpora such as ImageNet, GLUE, and AudioSet, olfaction lacks scaled equivalents and universally accepted benchmarks. This gap hinders progress and delays the achievement of intelligence milestones in artificial olfaction. Adaptive learning presents a critical path forward, enabling machine olfaction to evolve alongside advancements in computer vision, natural language processing, and auditory intelligence. In this survey, we explore why adaptive learning is essential for olfaction and highlight the instruments and theoretical foundations that uniquely position it to benefit from active sensing methodologies. We argue for the necessity of active and continuous learning over small datasets for attaining state-of-the-art performance in tasks such as classification, navigation, and general olfactory reasoning. Our review covers key components that facilitate adaptive learning, including multi-modal learning, swarm intelligence, game theory, neuromorphic computing, and uncertainty quantification. Through this survey, we aim to advance understanding of machine olfaction, chemical sensing techniques, and frameworks for active, continual machine learning. We hope to inspire interdisciplinary researchers to push the boundaries of olfactory robotics and drive progress in this crucial but underexplored domain.

***Keywords:*** machine olfaction, adaptive learning, continuous learning, robotics, artificial intelligence

Contents

Abstract 1
Contents 1
1 Introduction 2
1.1 Scope of the Survey 2
1.2 Machine Olfaction 2
1.3 Adaptive Learning 2
2 Theories of Olfaction 3
3 Olfaction Standardization 4
3.1 Data Standards 4
3.2 Benchmarks 4
4 Use Cases for Olfaction 5
5 Olfaction and Navigation 6
5.1 Sensing Techniques 6
5.2 Dynamic Environments & Ambiguity 7
5.3 Neuromorphic Computing 8
6 Plume Tracking & Olfactory Navigation 9
7 Multi-Modal Learning 10
7.1 Fusion vs Multi-Modal Learning 11
7.2 Grounding Modality 12
7.3 Multi-Kernel vs Single-Kernel Learning 12
7.4 Modality Encoders in Unified Models 13
7.5 Mixture of Experts Models 13
8 Adaptive Learning 14
8.1 Continuous & Federated Learning 14
8.2 Self-Adaptive Networks 15
8.3 Task vs Style vs Modality Transfer Learning 16
8.4 Multiple Methods, One Objective 17
8.5 Synthetic Data 17
8.6 Plasticity 17
8.7 E-Prop and Eligibility Traces 17
9 The Influence of Game Theory 18
9.1 Learning in an Unknown Environment 19
9.2 Frequentist vs Bayesian Statistics 20
9.3 How Behavior & Regret Influence Adaptability 21
9.4 Last-Iterate Divergence 21
9.5 Evidential Uncertainty 22
9.6 Application to Adaptive Learning 23
10 Swarm Intelligence 23
10.1 Reinforcement Learning 24
10.2 Multi-Agent Reinforcement Learning 24
10.3 Co-training and Co-regularization 24
10.4 Evolutionary Algorithms 25
10.5 Factoring the Search Space 26
10.6 Impact of Agent Design on Swarm Functionality 26
10.7 Expected vs Maximum Rewards 27
10.8 Emergent Behaviors in Swarms 28
10.9 Diffusion & Diffusion Policy 29
11 Environment & Simulation 29
11.1 "Sim2Real" Gap 30
11.2 Environment Design 30
12 Alignment & Ethical Considerations 31
13 Conclusion 33
References 33

# 1 Introduction

The field of robotics has gathered a lot of momentum over the last couple of decades partially due to the advances in many domains of AI. Humanoid robotics in particular have seen a surge of interest as of late. Companies developing these robots are attempting to replicate the five senses of humans artificially. Vision and hearing are the two highest-bandwidth modalities that humans sense with, so they deserve a high degree of research attention in order to replicate artificially. Computer vision and natural language processing have provided benchmark performance on sight and sound, with a significant part of the last few decades being dedicated to moving the performance in these fields forward to super-human level. Artificial touch has been thoroughly addressed by haptic feedback mechanisms and mechano-sensory actuators. Taste, while still a large contributor to human level intelligence, is arguably the least important sense because taste is largely involved in the method in which humans acquire energy and fuel, with most robotics acquiring their energy from electrical power. Therein leaves one final sense to be explored - the sense of smell.

## 1.1 Scope of the Survey

Artificial smell is commonly referred to as machine olfaction. This survey covers machine olfaction and adaptive learning. Olfaction is uniquely suited to benefit from adaptive learning due to the lack of standardization, large scale datasets, and proliferate community. We discuss various methods of learning and focus on those particularly fitting for robotics, active sensing, and embodied learning among stimuli from the environment. Reinforcement learning, multimodal learning, and continuous learning are machine learning approaches that fit these criteria. Reinforcement learning tightly couples with game theory, and game theoretic principles become important in multi-agent settings. As artificial intelligence becomes more capable, robotics will inherently become multimodal. Multimodal learning can uniquely benefit machine olfaction because vision, language, and audio can act as priors to inform olfactory posteriors.

This work takes the intersection of surveys in multimodal learning [13] [204], adaptive learning [201] [112], swarm intelligence [92], and machine olfaction [35] [48] [139] and contextualizes them with modern research to answer the singular question: *How can machines rapidly learn the sense of smell?*

## 1.2 Machine Olfaction

Research in machine olfaction is largely asymmetrical to the magnitude of research performed in computer vision and natural language processing, but the advancement of olfaction science should be given due importance.

Before vision or auditory capabilities, the most primitive forms of navigation in biological organisms arose from olfactory tracking of odor plumes and pheromone trails to a food source. Odor plumes are dynamic, change directions with wind shifts, and are highly subject to environmental constraints such as temperature and relative humidity. The strength of the plume also slowly decays with time as air equalizes, making it difficult to identify the tail of the odor plume. While some organisms may visually locate the source of an odor over short distances, they must rely entirely on olfactory tracking when the source is not yet visible, making scent-based navigation via plume tracking a very challenging control problem.

Successful plume tracking is dependent on appropriately diagnosing uncertainty, and we focus on demonstrating how adaptive learning helps with constructing a probability distribution that can be used to locate the plume source in machine olfaction tasks. Evolutionary techniques such as ant colony optimization (ACO) [60] and particle swarm optimization (PSO) [103] were originally inspired by insects that collaboratively perform pheromone tracking through scent plumes and these scenarios prove to be excellent models for our work.

## 1.3 Adaptive Learning

Conventional machine learning methods require thousands or even millions of training samples per class in order to accomplish a target task. The typical process of learning occurs via training over this dataset, validating over a separate holdout set, and then deploying of the model to the application. In many scenarios, this model is not ever trained again, which can lead to performance degradation upon data drift. This limits the applications for which machine learning can be used.

There has been a plethora of research performed over the last few decades of different ways to optimize artificial intelligence models in learning more over "big data". In proportion, artificial intelligence models have grown in size such that big models and big data elicit big results. With the growing popularity of large language models (LLMs), datasets on the order of petabytes are fed in for training, and these models are so large and training so cumbersome that it is very expensive to construct them. Engineers and researchers performing computer vision (CV) and natural language processing (NLP) tasks are privileged in the fact that terabytes data and benchmarks exist for download online. CV and NLP knowledge is also very prolific among the machine learning community, making it more likely to advance the state of the art and faster to develop solutions. In cases where samples are rare (e.g. certain medical applications), one may not always have the luxury of large datasets. Such is the case with machine olfaction. To achieve the same state-of-the-art in olfaction as in vision and audio, continuous updates of priors over existing data must take place. In the absence of large training

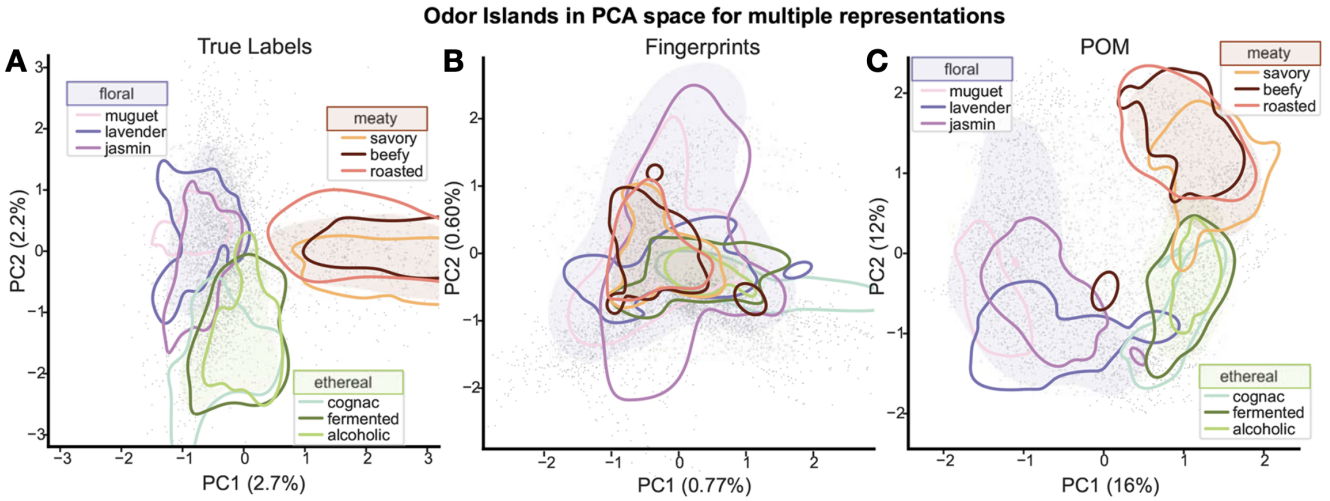


**Figure 1.** Odourants plotted by the first and second principal components according to their perceptual labels at (A), structural fingerprints at (B), and *Principal Odor Map (POM)* coordinates at (C). Areas dense with molecules having the broad category labels floral, meaty, or alcoholic are shaded; areas dense with narrow category labels are outlined. The POM recapitulates the true perceptual map, but the fingerprint map does not; note that only relative (not absolute) coordinates matter. See [110] for further details. Figure and caption adapted from Lee, et al. in [110] under the Creative Commons License.

sets and credible simulations, adaptive learning is the only method by which a machine can acquire intelligence.

## 2 Theories of Olfaction

The position paper from [74] highlights why many issues in olfaction stem from the lack of a universal data standard and the contributing reasons for this. Olfaction stands apart from other sensory modalities in that it admits multiple physical mechanisms for detecting and discriminating odorants. Depending on the underlying sensor architecture, these mechanisms can span distinct physical domains. Certain optical sensors infer molecular presence by monitoring shifts in specific wavelength bands as odorant molecules perturb the refractive or absorptive properties of a medium. Electrochemical detectors, in contrast, measure current or voltage differentials induced by redox reactions or ionic transport, typically across an electrolyte or semipermeable membrane. Other approaches operate at the quantum scale, detecting odorants by measuring molecular vibrational spectra—essentially "hearing" the unique frequency signatures associated with intramolecular bond dynamics.

The former two sensing paradigms are most often associated with the *Shape Theory of Olfaction (STO)*, which posits that molecular geometry, size, and surface properties govern receptor binding and olfactory perception [19, 165, 194]. The latter vibrational approach corresponds to the *Vibrational Theory of Olfaction (VTO)*, an alternative hypothesis asserting that molecular scent is predominantly determined by quantized vibrational modes.

Originally proposed by C.G. Dyson in 1928 and further elaborated in 1938 [62, 125], the VTO faced early skepticism due to Raman spectroscopy data that failed to correlate vibrational spectra with perceived odor. However, in 2001, biophysicist Luca Turin reignited interest in the theory, proposing that inelastic electron tunneling might underlie odorant discrimination in mammalian receptors [180]. This theory has since been expanded to include variants such as phonon-assisted tunneling, as discussed by Brookes et al. [30]. Despite these developments, the theory remains controversial; Block et al. [22] have published extensive critiques challenging both its biophysical plausibility and experimental reproducibility.

Crucially, both STO and VTO enjoy empirical support from distinct experimental paradigms, yet neither provides a complete or unified account of olfactory phenomena. Notably, both theories fail to explain why certain odorants exhibit concentration-dependent perceptual shifts—a phenomenon well recognized by perfumers but still lacking mechanistic elucidation [26]. As Stevens remarked in 1951, “it is very probable that no one physical property alone is involved in the physical nature of the adequate stimulus” [174]. Thus, despite a growing body of work, a comprehensive and experimentally validated theory of olfaction remains elusive. No current sensor modality, whether biologically inspired or physically engineered, offers a complete framework for odor characterization.

This epistemological gap raises a practical question: how can one define a data standard for a sensory modality whose fundamental operating principles remain unresolved? We

argue that this lack of consensus is not a hindrance, but rather a fertile opportunity for the artificial intelligence and machine olfaction communities. Precedents from other domains are instructive. The development of JPEG and PNG standards did not require a complete understanding of human visual cognition—though progress in neuroscience has certainly informed modern computer vision. Likewise, aviation standards do not demand biomimicry of avian flight; yet regulatory frameworks for airworthiness have enabled safe and scalable deployment.

By analogy, the formulation of olfactory data standards can proceed alongside, rather than in wait of, a final theory of olfaction. Recognition of both STO and VTO perspectives should inform the design of benchmark tasks, sensor calibration protocols, and data representation formats. But standardization and theoretical resolution need not proceed sequentially; they can—and should—co-evolve. We point interested readers to the work of [74] for a more comprehensive analysis on the points presented in this section.

## 3 Olfaction Standardization

### 3.1 Data Standards

In contrast to other sensory modalities, olfaction lacks a universally accepted data standard. Visual information is typically stored in well-defined formats such as PNG and JPEG (for images) and MP4 (for videos), which facilitate uniform processing and widespread data sharing. Audio data benefits from standardized representations like WAV files, while speech can be transcribed into words that serve as a natural language standard. These common formats have been instrumental in driving rapid progress in machine learning applications for vision, audio, and speech processing.

Olfactory data, however, has no such analogue. There is no agreed-upon digital file format or encoding scheme that comprehensively captures the rich, multidimensional nature of odor perception. This absence of standardization has led to a paucity of large, curated datasets, which in turn hampers the development of robust machine learning models for odor prediction and synthesis.

Recent research has sought to bridge this gap through innovative approaches. For instance, the work by Lee, Wiltschko and colleagues introduces a *Principal Odor Map* (POM) that represents odors as nodes within a graph structure, where inter-node distances reflect perceptual similarities between odorants [110]. This approach offers a promising route to "digitize" odor by creating a high-dimensional mapping analogous to RGB for color. Their training data is a combination of the GoodScent [46] and LeffingWell [11] datasets, which give human-evaluated aromas for several chemical compounds. This Principal Odor Map deduced from PCA space can be seen in Figure 1. Complementary data from research on the olfactory perception of structurally diverse chemicals was produced by Keller in [102].

A large training corpus of 18,000 time series measurements over 72 gas sensors was compiled by [186]; however, later work from Dennler, et al. in [55] revealed that the lack of certain experimental controls during data accumulation could invalidate the prudence of the data.

In parallel, research led by Michael Schmuker and Nik Dennler has modeled olfactory responses as time series, capturing the dynamic, temporal nature of odor signals through neuromorphic circuits [54]. These time series representations mimic the spiking behavior observed in biological olfaction, yet they remain isolated demonstrations rather than components of an integrated, universal standard.

The proliferation of the attention mechanism and the transformer architecture by Vaswani, et al. [185] has popularized the idea of breaking data up into small chunks called "tokens". In text, each token represents a word or word part; in vision, each token represents a group of pixels. Transformers train to understand patterns within these tokens and contextualize them to larger patterns within the data sample. Analogously, one can envision scent being broken up into tokens where each token is a representation of a molecule at a specific temperature, or a group of molecules belonging to a specific aroma. While this has yet to be exercised, the prolific use of transformers in modern machine learning begs the idea to be proven.

The current fragmentation in olfactory data representations — not only between graph-based and time series approaches but also across diverse experimental protocols — presents a significant barrier to data aggregation and model generalization. Without a common standard, datasets remain small, heterogeneous, and difficult to compare, which slows progress in both fundamental olfactory research and its practical applications.

Moving forward, the establishment of a standardized data format for olfaction could catalyze advances similar to those seen in vision and speech processing. Such a standard would not only enable the integration of disparate datasets but also foster the development of more robust machine learning frameworks tailored to the complexities of odor perception.

### 3.2 Benchmarks

The lack of a data standard makes it difficult to define a set of benchmarks from which olfactory sensors should be measured. Language models are benchmarked against GLUE and ROUGE. Computer vision models are benchmarked against CIFAR and ImageNet datasets. Larger models such as LLMs are measured against various benchmarks for coding, mathematical reasoning, creative writing, visual question and answering (VQA), and deep research [113]. Many of these benchmarks can be viewed on the popular HuggingFace leaderboard [42] that give a live ranking of top performers.

While incorporating olfaction into measures for mathematical reasoning could grant little advantage, it could strongly influence VQA tasks and instruction finetuning over

chemistry and agricultural tasks. One can imagine the advent of olfaction-visual instruction finetuning and olfaction-visual question and answering (OVQA) benchmarks as a result.

The Open X-Embodiment dataset [45] is a multimodal dataset designed for vision-language robotics applications. It is a collaboration between 21 institutions collected over 160k tasks from 22 different robots. Such a collaboration is extremely monumental for the realm of robotics and the addition of olfaction at this scale could enable a strong series of measurable benchmarks that assess the incremental advantage received by adding olfaction to a task. This also calls creates further opportunities to improve multimodal model alignment, which is discussed more in section 12.

## 4 Use Cases for Olfaction

Many applications of olfactory sensors exist. One of the most commonly recognized uses is the smoke detector, a chemical sensor that detects high concentrations of carbon monoxide. The below summaries give a brief overview of some common applications within olfactory sensing with a special focus on scent-based navigation in Section 6.

**Robotics** Companies like Boston Dynamics [25], Tesla [195], Figure [65], and Clone Robotics [44] are developing humanoid robots with ambitions to deploy them at scale—in homes and factories—over the next decade. These robots are typically equipped with cameras and lidar for vision, microphones and speakers for audio, and embedded language models to facilitate human-like communication. Yet, one critical sensory modality remains absent: olfaction. Incorporating olfactory sensors could dramatically expand the functionality of these systems, enabling robots to detect hazardous chemicals, perform breath-based health monitoring, and navigate environments by scent cues [34, 35, 61].

**Agriculture** In agriculture, olfactory sensors are valuable tools for monitoring soil health through the detection of volatile organic compounds (VOCs) emitted by microbial activity [40, 57]. They also offer early disease detection by sensing VOC markers released by infected plants, allowing farmers to act preventively and reduce crop losses—while minimizing reliance on chemical treatments. Recent work by Barhoum et al. [16] outlines the engineering challenges and opportunities in designing olfactory systems that function reliably in rugged agricultural environments.

**Food and Beverage Industry** Aroma is a crucial indicator of quality and authenticity in wine and other consumables. Olfactory sensors can detect nuanced scent profiles that may reveal contamination or counterfeiting, thus protecting both consumers and brand integrity. Aryballe's work with Mach-Zehnder interferometer-based sensors illustrates the potential of artificial olfaction in beverage authentication [10]. Longin et al. [119] have shown that bread aroma analysis can predict shelf life, while Wang et al. [188] apply similar techniques to assess pork freshness. Together, these efforts demonstrate scalable frameworks for broader food quality assurance.

**Indoor Air Quality Monitoring** In residential, commercial, and public spaces, olfactory sensors enable continuous monitoring of indoor air quality. They can detect pollutants such as VOCs, allergens, mold, and gases infiltrating from the outdoors [90], often at concentrations below human detection thresholds. Humans themselves emit VOC signatures [96], and olfactory systems may learn to associate chemical profiles with specific occupants. Much like smoke detectors are now ubiquitous, advanced olfactory sensors could become a standard feature in indoor environments—enhancing safety, comfort, and public health. Opportunities for innovation in this space are still substantial [16].

**Cosmetics & Perfumery** In the cosmetics and fragrance industries, olfactory sensors are aiding product formulation and quality control for items like lotions, creams, and deodorants. Companies such as Osmo AI [142, 143] are pioneering the intersection of artificial intelligence and olfaction to both generate novel products and maintain olfactory consistency in existing ones. These sensors help ensure that scent profiles match consumer expectations and that no unwanted odors compromise product integrity.

**Energetics and Explosives Detection** Olfaction is also being explored for detecting explosives and hazardous substances, in both military and civilian domains. Canine units remain the standard [52], but artificial alternatives are emerging. Surveys by Bobrovnikov [24] and others [148] catalog the state of sensor-based explosive detection. Wasilewski et al. [191] propose hybrid systems that combine biological and artificial approaches. Drone-based olfaction for locating explosives is also under investigation [35], opening new pathways for automated surveillance and protection. Research from [120] emphasize the difficulty in working with such a use case with small datasets - another point to underscore the importance of adaptive learning in olfaction.

**Personalized Medicine** In healthcare, olfactory sensors offer noninvasive means to tailor treatments through breath and fluid analysis, enabling personalized diagnostics based on metabolic signatures. These sensors can detect disease-specific biomarkers—for conditions such as pneumonia [15, 100] and lung cancer [6, 146]—and track how individuals respond to treatments in real time. This paves the way for more targeted and responsive medical interventions.

**Automotive Industry** The automotive sector is exploring olfactory sensors to enhance in-cabin air quality and monitor for hazardous odors [9, 179]. These sensors can elevate passenger comfort by neutralizing unpleasant smells and alerting to contaminants. Additionally, they serve a role in manufacturing, where olfactory monitoring helps ensure that car interiors meet sensory quality standards and can detect early signs of material degradation, such as paint or corrosion issues [152].

## 5 Olfaction and Navigation

Organisms navigating through olfactory methods use a variety of behaviors to locate the plume source. Two prominent behaviors are called casting and surging. Casting is the act of scanning from side to side in gradually broader strokes. Surging is the act of rapidly moving upwind of the plume. In nature, insects and ants will typically cast until they acquire the plume and surge once they detect it; if they lose the plume, they will repeat the process until the source is located. Ants use a method of detection called tropotaxis with symmetrically placed receptors in the antennae.

Chemotaxis is the process of continuously sampling the environment with a single receptor and moving in the direction of the strongest gradient. Bacterium navigate through a process called orthokinesis which compares the temporal change in stimulus intensity and makes changes in movement analogous to such intensity. The research of [169] and [168] leverage a combination of both chemotaxis and orthokinesis because their methodology samples the environment with a single sensor and measures the change in magnitude between time steps, accelerating (or surging) in proportion to the magnitude change. This mechanism gives natural inclination to use various temporal difference reinforcement learning (RL) agents with which swarms can be constructed. Olfactory-based scenarios like plume tracking are an excellent and unbiased way of exemplifying adaptive learning because they require quantifying a high degree of uncertainty, characterizing the environment, multi-modal learning, and multi-agent cooperation, effectively consolidating each of the aforementioned principles.

### 5.1 Sensing Techniques

To properly qualify the application of adaptive learning techniques in the context of olfactory-based navigation, it is helpful to establish some of the limitations of existing olfactory sensors and why their limitations impose certain restrictions on learning. Cameras and microphones can rapidly sample visual and audio data, respectively, hundreds of times per second. The technology is mature enough that only a small temporal assessment is needed to identify patterns in the respective data mediums. Olfactory sensors are inherently slow, the fastest sensors taking among seconds and the slowest taking among hours to properly respond to a stimulus. The environmental conditions for olfactory stimuli change much faster. This makes dynamic responses for navigation tasks difficult because agents typically need to respond quickly due to changes within the environment. Crimaldi, et al. in [49] highlight the paradox with these temporal mismatches well.

The aim of their work is to communicate the evidence of high temporal resolution present in olfaction data. The slow sensing mechanisms are largely fixed by physics, but the environment data can be adjusted to accommodate the sampling frequency of the sensor. However, [49] suggests an alternative route of "reformatting the spatiotemporal structure of an odor field into temporal fluctuations registered by a sensor, with both a finer spatial structure or faster relative motion (flow-to-sensor) leading to higher-frequency fluctuations." Put simply, one can reformat the data received by the sensor into spatiotemporal data that maximizes information over a smaller sampling rate. They suggest fusing other sensing modalities together to maximize the spatiotemporal picture at each timestep and employing multi-modal learning techniques on top of this. The previously mentioned work of ImageBind in [80] could be an excellent candidate here: instead of encoding all input into visual data, one could encode all input into spatiotemporal data.

**5.1.1 Gas Chromatography Mass Spectrometry** Gas Chromatography-Mass Spectrometry (GC-MS) is a powerful analytical technique used to identify and quantify compounds in complex mixtures. The process involves two main steps: Gas Chromatography (GC) and Mass Spectrometry (MS). GC-MS is considered the gold standard for chemical detection due to its high sensitivity and specificity, versatility in analyzing a wide range of compounds, quantitative accuracy, and reliability. It can detect compounds at very low concentrations and provides precise quantitative data, making it suitable for various applications, including environmental analysis, forensic science, food safety, and pharmaceuticals. The combination of GC's separation capabilities with MS's detailed molecular analysis ensures consistent and reproducible results, making GC-MS an indispensable tool for chemical detection and analysis.

**5.1.2 Metal Oxide Sensors** Metal oxide sensors are widely used for gas detection due to their simplicity and cost effectiveness, consisting of a sensing layer made from metal oxides like tin oxide. They operate by detecting changes in electrical resistance when target gases interact with the metal oxide surface, making them effective for gases such as carbon monoxide and methane. These sensors are inexpensive and have a fast response time but can be sensitive to humidity and temperature changes.

**5.1.3 Photoacoustic Sensors** Photoacoustic sensors leverage the photoacoustic effect, where absorbed light energy is converted into sound waves, allowing for the detection of trace gases. When a gas absorbs light from a modulated laser, it generates acoustic waves that are detected and analyzed, making these sensors suitable for environmental monitoring and industrial applications. They provide high sensitivity and are less affected by temperature variations but can be complex and costly.

**5.1.4 Nondispersive Infrared Sensors** Non-dispersive infrared sensors (NDIR) utilize infrared absorption spectroscopy to detect gases by measuring the absorption of infrared light at specific wavelengths. They consist of an

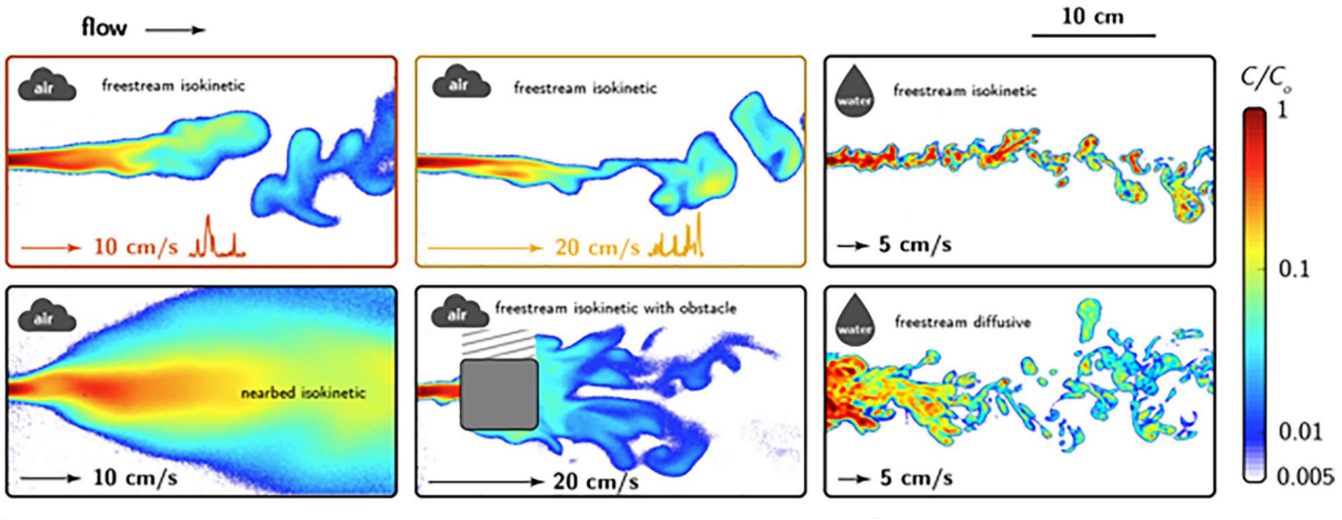


**Figure 2.** Odor landscapes and temporal reformatting of spatiotemporal structure. Normalized instantaneous odor concentration fields measured by planar laser-induced fluorescence illustrate diverse odor landscapes in air (left & middle columns) and water (right column) for varying release conditions and flow speeds. Cross-hatching signifies a data gap from laser shadowing behind the obstacle. Figure and caption adapted from [49] under the Creative Commons License.

infrared light source, a gas sample chamber, and a detector, making them effective for measuring gases like carbon dioxide and methane. NDIR sensors offer high specificity and sensitivity, are stable over time, but can be more expensive and require periodic calibration.

**5.1.5 Electrochemical Sensors** Electrochemical sensors are used for detecting toxic and combustible gases, consisting of an electrode system immersed in an electrolyte. They operate by generating a current proportional to the gas concentration when the target gas undergoes a redox reaction at the electrode. These sensors are highly sensitive and relatively inexpensive but may require regular calibration and can be affected by other gases. Work from [147] [82] and [15] denote several applications with electrochemical sensors.

**5.1.6 Conductive Polymer Sensors** Polymer sensors utilize conductive polymers that change their electrical properties in response to specific chemicals, operating on the principle that gas absorption leads to changes in conductivity or resistance. They are often used for detecting volatile organic compounds and can be tailored for specific applications, being lightweight and low-cost. However, they may have lower sensitivity compared to other sensor types and can be influenced by humidity and temperature changes.

**5.1.7 Colorimetric Sensors** Colorimetric sensors detect and quantify chemical substances based on color changes resulting from chemical reactions. These sensors typically consist of a sensing material that interacts with specific analytes, leading to a visible color change that can be measured and correlated to the concentration of the target substance. They can be designed to respond to specific gases or compounds, providing a visual indication of their presence. The degree of color change can be quantitatively analyzed, allowing for the measurement of concentration levels of various substances. Colorimetric sensors can be coupled with cameras or spectrometers to automate the detection process and enhance data analysis.

### 5.2 Dynamic Environments & Ambiguity

Even through the use of fast and reliable sensing techniques, navigation via olfactory signal is still a very challenging task. Singh, et al. in [168] and [169] demonstrate a method using reinforcement learning to train a recurrent neural network that can maintain temporal information about changing plume dynamics. They maintain the protocols of [79] and [59] in keeping the agent simple, computationally efficient, and adaptive. They employ similar principles that complement active, continuous, and lightweight learning.

Crimaldi discusses how insects use multimodal principles through fixational eye movements while tracking plumes. Locking onto a source of smell visually allows the insect to discern through ambiguity imposed by air dynamics. Insects actively adjust their antennae to scan for odors as they are tracking. This effects the neural odor representation within their brain. Mammals sample odors in the environment through rapid sniffing at 2-15 Hz, pulling the target odorants into contact with the olfactory bulb in the process. While these techniques are used to maximize the signal within the olfactory sensor, they also introduce different dynamics around the immediate local environment surrounding the sensor. This introduces more ambiguity into

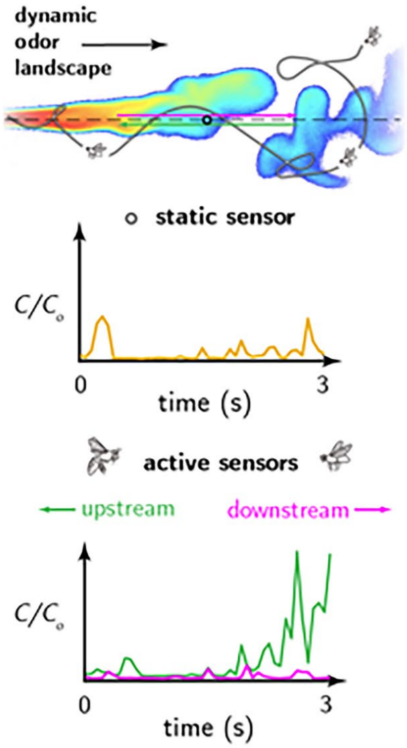


**Figure 3.** Moving through odor landscapes is an active sensing modality where the information content of the signal is modified by sensor kinematics (top panel, black line shows a hypothetical trajectory). This is seen in concentration time-series from one static sensor (middle panel, black circle in top panel), and two active sensors (bottom panel, green & magenta arrows in top panel) moving upstream (downstream) through the same plume along straight trajectories (arrows, upper panel) on the mean plume centerline (dashed black line, upper panel) at 5 cm/s absolute velocity. All sensors have the same mean position over their 3-second trajectories (black circle symbol, top panel). Figure and caption adapted from [49] under the Creative Commons License.

the sensor sampling and further magnifies the difficulty behind olfaction use cases. Many aspects of olfactory tracking can be modeled by navigation algorithms that consolidate a simple state machine with continuous updating of heading and orientation governed by instantaneous detection of the odor concentration [49]. This makes the problem of olfactory tracking seem simple, especially when multiple modalities are involved. Where the complications arise is through the modeling of the odors themselves and the continuous motion (e.g. surging and casting) of the insect. Aerodynamic modeling is inherently difficult, highly dimensional, and has many degrees of freedom. From [49], "Hopfield, 1991 suggested that odors emanating from spatially separate sources usually generate distinct spatio-temporal distribu- tions, whereas co-located odor sources will result in coincident odor encounters." Knowing these proxies about aerodynamic behavior allows us to build higher degrees of uncertainty quantification into the simulation environment.

Drift among olfaction sensors is also a common problem. Each sensing type drifts with its own variance and different environmental variables can significantly influence this. Metal oxide sensors, for example, have a warmup duration that must be completed before the sensors can be reliably used for gas measurement. It can be difficult to discern sensor drift between true signal after this warm-up period occurs, because environmental factors, sensor saturation, and time since last use can all influence the true duration of the drift period needed for adequate sensor reset. Dennler, et al. in [55] demonstrate different remedies for detecting drift and how to compensate for it through methods like bout detection [162].

The dynamism in olfaction scenarios make acquiring training data difficult, and acquiring matching multi-modal sensors even more so. As a result, adaptive learning through multi-modal and multi-agent methods becomes not only helpful, but paramount for success.

### 5.3 Neuromorphic Computing

Neuromorphic computing is an approach to designing computer systems inspired by the structure and function of the human brain. Instead of relying on a centralized clock and separate memory and processing units, neuromorphic systems use networks of spiking neurons that process information asynchronously and in parallel based on events. This design mimics how biological neurons communicate through discrete events (spikes), leading to highly energy-efficient, real-time processing, particularly well-suited for tasks in active and continuous machine learning. Most neuromorphic architectures co-locate memory and processing together, reducing data movement and energy consumption [154].

Neuromorphic architectures naturally support massive parallelism. This can be advantageous for applications in artificial intelligence and sensory processing, where many small, concurrent operations are needed. Many neuromorphic systems use spiking neural networks that are well-suited for processing noisy or unstructured data, potentially offering robustness in areas like autonomous systems or robotics. Leveraging insights from the human brain could lead to more adaptive, resilient, and efficient computing systems that better mimic human cognitive functions, opening doors to new types of machine intelligence.

While neuromorphic computing is still largely in its infancy, some real world applications are being developed. Work from [95] specifically demonstrate how neuromorphic architectures naturally exhibit adaptive learning through the construction of a real circuit. Research from Hajizada, et al. in [85] has shown one of the most promising real-world developments in neuromoprhic computing with their Loihi

chip. The Intel Loihi chip is designed to mimic key aspects of biological neural processing.

Adaptive learning naturally inherits from neuromorphic computing, and machine olfaction can substantially benefit from its event-driven architecture. When it comes to olfaction, many challenges revolve around processing complex, noisy, and time-varying chemical signals. Olfaction sensors detect volatile compounds that might appear sporadically. By encoding sensor outputs as asynchronous events–much like spikes–one can process only the meaningful changes rather than continuously sampling, which leads to lower power consumption and faster response times. The spiking neural networks on a neuromorphic chip are inherently good at capturing the temporal dynamics of signals. This is a natural fit for olfaction, where the pattern and timing of chemical concentrations can be crucial for identifying different odors. With on-chip learning, a neuromorphic system can adapt to variations in environmental conditions, sensor drift, or new odor profiles.

As shown by the work of Dennler, et al. in [56] neuromorphic computing highly complements applications for olfaction. Yet this idea is still largely in experimental or early commercial stages. Research by [107] delineate challenges and opportunities of neuromorphic computing at scale. Issues like manufacturing consistency, scalability, and durability under varied workloads need to be addressed before they can be considered a reliable replacement for mature technologies. The tools, frameworks, and programming models for neuromorphic computing are not as well-developed as those for conventional architectures. This makes it challenging to design, implement, and optimize applications—hindering widespread adoption. Many of today's algorithms, particularly in machine learning, are designed for conventional hardware. Adapting or rethinking these algorithms to leverage neuromorphic principles may require significant innovation and re-engineering. While neuromorphic systems may excel in specific domains (like sensory processing or edge computing), it's unclear whether they can handle the broad range of tasks managed by general-purpose von Neumann systems. A hybrid approach might emerge, where neuromorphic processors complement rather than replace conventional architectures.

## 6 Plume Tracking & Olfactory Navigation

The machine olfaction task that can most benefit from adaptive learning is plume tracking, or scent-based navigation. Plume tracking not only involves the act of localizing an odour to its source, but in identifying the specific compound to track and modeling the plume itself. Accurate plume tracking requires the robot to create a model of the plume and understand how the plume evolves in accordance to varying environmental conditions like wind, temperature, and humidity. The signal magnitude for many chemical compounds change as they are heated, pressurized, and mixed with other chemicals in the air. This means that the robot must intuit how the target compound is evolving as it navigates. For example, for a robot trained to locate the source of the compound 2,4-Dinitrotoluene (a primary component of explosives), the signal for such compound looks vastly different at 0°C and 2000 meter altitude than it does at 40°C and 100 meter altitude. Thus, effective plume navigation must balance an ego-centric model from the robot and a world model of the changing plume. In some instances, multi-hypothesis tracking and extended Kalman filtering can prove to be effective. However, in extreme environments, these Kalman filters may need to be fused with neural models like those from HybridTrack in [17] maybe needed. This is also partially remedied by the techniques defined in [71]. In their work, they define analytical techniques that are helpful for calibrating various olfactory sensors. Their research builds off that from [72] where they demonstrate a method called *Olfactory Inertial Odometry (OIO)* that fuses inertial data with the olfactory signal of chemical sensors to facilitate scent-based navigation.

Research from [49] discuss the dynamics of plume tracking. Computational fluid dynamics packages become beneficial in modeling how plumes will behave for various environments and compounds. Even the effect that the robot itself has on the plume must be considered: The rotors from a quadcopter can significantly distort the plume as it is being tracked. This warrants the robot designer to consider careful placement of the olfactory sensors to maximize the signal. In the demonstrations of [34] and [61], both show that placing the olfactory sensor at an elevated point in the middle top of the quadcopter is sufficient to track ethanol. On the other hand, the antennae of moths extend far passed their head in order to escape the beat of their wings, so that the olfactory sensors of the antennae may sense the undisturbed air.

Other robots such as humanoids may not have as strong of an influence on a plume being tracked since their movements are more graceful and induce less turbulence than propellers. However, dense chemical compounds that lie low to the ground may difficult for tall humanoids to detect. Examples from [35] discuss how quadcopter rotor wash reverberating off the ground strongly distorts the olfactory signal when tracking a plume close to the ground. This is in contrast to ants that perform pheremone tracking by laying chemicals close to the drone for the rest of the colony to localize.

Adaptive learning is a very appropriate methodology from which plume tracking can benefit due to the volatility of plumes and the influence even subtle aerodynamic changes have on chemistry, and Crimaldi et, al. allude to this in their 2022 work. Methods from Singh, et al. in [169] leverage many of their assumptions to train a recurrent neural network (RNN) through reinforcement learning. An overview of their assumptions for plume modeling is displayed in Figure 4.

Figure 5 illustrates how the trained RNN handles changes in plume direction and differing plume diffusion rates. Singh, et al. demonstrate their methods in simulation, but an opportunity exists to prove the practicality of their work in reality through experiments similar to [34] and [61].

Though only a singular problem in machine olfaction, plume tracking engages several different disciplines and no one unified framework has been proposed that effectively enables robots to navigate by scent. Adaptive learning is key to solving olfactory navigation and doing so will largely be driven by a confluence of sufficient data acquisition over a defined olfaction standard, a variety of olfaction sensors, and multimodal learning.

## 7 Multi-Modal Learning

Data and sensing can be classified into different modalities. Each modality is a quantitative format for which different machine learning techniques can be used. Vision, hearing, and speech are all different sensing modalities and they are all represented by PNG/JPEG data (for images), WAV/MP3 data, and text data respectively. As previously mentioned, there is not a community consensus on how olfaction data should be represented. In most cases, olfaction data is represented in the form of time series data or spectrograms. This survey shall focus on the general field of olfaction, discussing the many different sensing techniques and data forms available. Some of the more compelling demonstrations of olfaction are not where olfactory sensors are used in isolation, but used in conjunction with other sensing modalities such as in computer vision for scent-augmented navigation.

Multi-modal learning is the process of using data inputs of multiple types in order to construct a model of approximation. These modalities can be image data from cameras, audio data from microphones, chemical data from olfaction sensors, or any other kind of data medium. One approach to multi-modal learning is to distinctly learn separate signals for each modality through separate models. This works well in theory, but can become highly computationally expensive as many modalities are added.

An alternative method of multi-modal learning is called *modality binding*. In their work [80], they discuss *ImageBind*, a model that fuses these data streams together into a single "embeddings space" that facilitates conditional pattern recognition in higher dimensions. Multi-modal learning then occurs by conditioning all other modalities with respect to one "grounding modality". They conduct experiments over six different modalities: images, thermal, motion (inertial measurement unit), audio, text, and depth data. All of these modalities are bound to the image modality, hence the name "ImageBind". One benefit of having a single embeddings space for all modalities is the break of dependence among all modalities contributing to multi-modal learning. Multi-modal models can be developed in such a way that all six modalities must be needed as input in order for an accurate response to occur; if one modality is not present in the input, the model exhibits high uncertainty or breaks. With ImageBind, the authors remove this crux through the creation of the common embeddings space. One modality, all six modalities, or an observation with any permutation therein can be fed into the model and still deliver a competent result. This requires identical encoders among all data modalities so that an entire observation can be encoded into the same latent space and therefore be normalized for training. In their work, [80] demonstrate that modality pairs are "bound" together to create this embeddings space. They define $I$ as the grounding modality and $M$ as another modality - any of the other five. These modality pairs are encoded together, effectively binding them in the process. These encoders are optimized through an InfoNCE loss, first proposed by [181]. The equation for this loss is shown below.

$$L_{I,M} = -\log \frac{\exp\left(q_i^\top k_i/\tau\right)}{\exp\left(q_i^\top k_i/\tau\right) + \sum_{j\neq i} \exp\left(q_i^\top k_j/\tau\right)} \tag{1}$$

where $\tau$ denotes a scalar temperature that controls the smoothness of the softmax distribution and $j$ defines "negatives", or unrelated observations. Each example $j \neq i$ in the mini-batch is presumed to be a negative. The loss makes the embeddings $q_i$ and $k_i$ closer in Euclidean distance in joint embedding space. This consequently aligns $I$ with $M$ conveniently for joint learning. Equation (1) significantly increases the efficiency of jointly learning the embedding space.

Other methods such as Prototypical Networks in [170] make use of Kullback-Leibler divergence as another helpful routine for use in multi-modal learning. The equation for this is shown below:

$$D_{KL}(P||Q) = \sum_{x\in X} P(x) \log \frac{P(x)}{Q(x)} \tag{2}$$

The "KL" divergence as it is commonly referred to measures the difference between two distributions. In multi-modal learning, each of these distributions can be thought of as characterizing a distribution around unimodal data. In high dimensional space, these distributions are not expected to align. However, with the use of the techniques in ImageBind, all modalities are encoded using a similar encoder and bound to a common reference modality. This enables the distributions for each modality to overlap, such that the KL-divergence is close to zero (zero meaning distributions are identical in shape, size, and spread). Coupling KL-divergence with the NCE loss could enable efficient adaptive learning with ImageBind. As Snell et al. suggest in [170], the use of KL-divergence enables out-of-distribution detection in prototypical networks. This concept can be extrapolated to ImageBind to enable active multi-modal learning, allowing agents to adapt to new environments and react to different

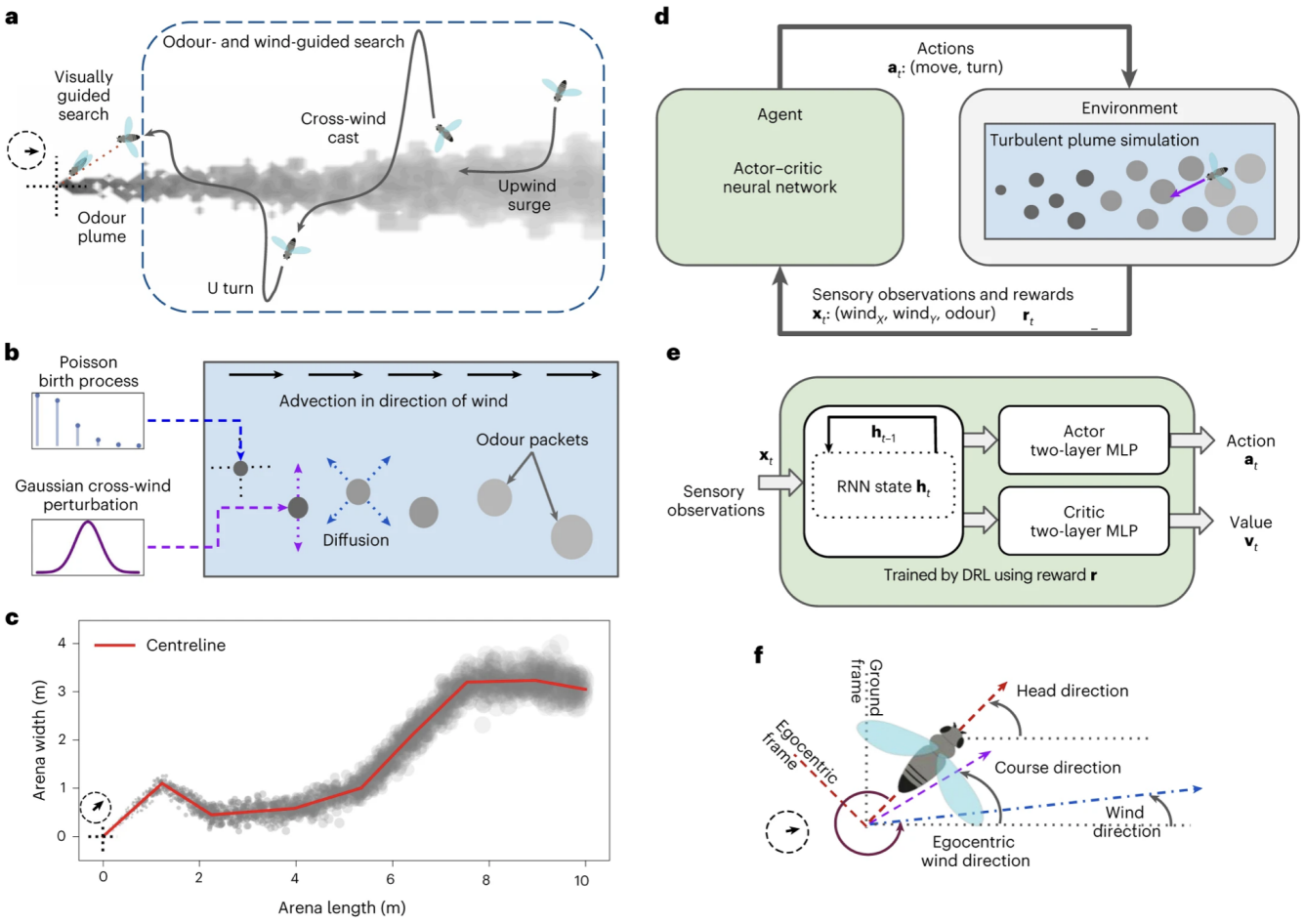


**Figure 4.** (a) A schematic of a flying insect performing a plume tracking task. (b) A plume simulation model showing stochastic emission of odour packets from a source carried by wind. Odour packets are subject to advection by wind, random cross-wind perturbation and radial diffusion. (c) An example of a plume simulation where the wind direction changed several times. (d A schematic of how the artificial agent interacts with the environment at each time step. The plume simulator model of the environment determines the sensory information *x* available to the agent and the rewards used in training. The agent navigates within the environment with actions *a*. (e) Agents are modeled as neural networks and trained through reinforcement learning. An RNN generates an internal state representation *h* from sensory observations, followed by parallel actor and critic heads that implement the agent's control policy and predict the state values, respectively. The actor and critic heads are two-layer, feedforward networks. (f) A schematic to illustrate an agent's head, course, and wind direction. Figure and caption are adapted from [169] under the Creative Commons License.

agent behaviors. The ability to detect and segregate out-of-distribution data using the methods defined above is crucial for complex olfactory tasks and critical for ensuring minimal computational complexity of edge algorithms on olfactory robots.

### 7.1 Fusion vs Multi-Modal Learning

Fusion and multi-modal learning are often referred to in the same context, many times as the same concept. Both concepts are similar in theory, but quite different in practice. Fusion is the step that occurs before multi-modal learning. In other words, all data modalities must be "fused" together in a way that normalizes all of the data and establishes a covariance model. A good example of this is the extended Kalman filter. Extended Kalman filters "extend" the concept of a single-variable Kalman filter by adding more variables into the filtering process. All of these variables are fused together to construct a covariance matrix that indicates how each input conditionally varies in the presence of other variables.

The joint embeddings space from ImageBind acts in a very similar way, but in much higher dimensionality. The embeddings space from ImageBind is analogous to the covariance matrix in an extended Kalman filter. Each variable (or data

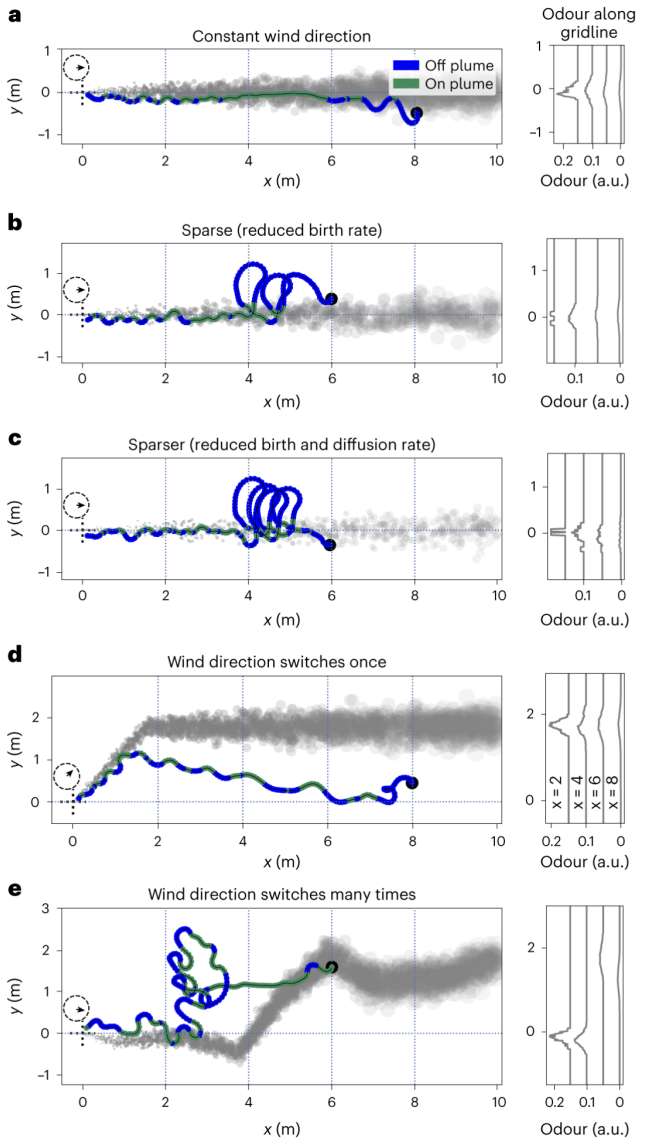


**Figure 5.** Snapshots of various odour plumes, denoted in grey, overlaid with learned agent trajectories. The trajectories are colored according to whether the agent was able to sense the presence (green) or absence (dark blue) of the odor. Trajectories start at the filled black circle and end at the odour source, indicated by dotted cross-hairs in the left-hand side of each panel. All examples use a 0.5 m/s wind. Figure and caption leveraged from the original work of [169] under the Creative Commons License.

modality) is put through an encoder before it is fused into the embeddings space. In reality, a single variable ingested by a Kalman filter is usually low in dimensionality (e.g. a scalar value) and complexity, whereas a single variable ingested by a multi-modal model like ImageBind is high in dimensionality (e.g. videos) and much more complex. Fusion is a part of the multi-modal learning process, but it is not the same as multi-modal learning.

### 7.2 Grounding Modality

The success of ImageBind in [80] shows how effective multi-modal learning can be by encoding all modalities into a common latent space. This can be achieved by selecting one modality and binding all of the others to this. In the case of ImageBind, image data was selected as the grounding modality to bind all others to. All other modalities were encoded into images in order to normalize the data into a common format and embedding space. In other words, from (1), $I$ always represents an image, and $M$ always represents one of either thermal, IMU, depth, video, or text data.

While this proves effective in learning, the difficulty in achieving this comes in ensuring pairs of images with each modality exist. Otherwise, the ImageBind technique cannot occur. This is not necessarily a fault with ImageBind, but a condition of multi-modal learning in general. Without paired data observations, controlled learning becomes difficult to achieve. In the case of ImageBind, the authors had to acquire observations for which all six modalities existed. This makes data acquisition very expensive and time consuming, but a necessary expense in order to build a high-fidelity model.

From a machine olfaction perspective, ImageBind could significantly help the work of [169] and [49]. Both of their works denote olfaction-only evaluations. Coupling another modality in their experimentation has the potential to significantly improve their work. Leveraging the methods from [80] and maintaining imaging as the grounding modality, olfaction data could be encoded into spectrograms in order to create a synergistic pair.

### 7.3 Multi-Kernel vs Single-Kernel Learning

Methods involving multi-modal learning with multiple kernels (or models) are an extension of the work from [81] where they show support vector machines can allow for the use of different kernels for different modalities or views of the data [1]. Kernels can be seen as similar functions or models between data modalities, each one allowing for better fusion of heterogeneous data [204].

Multi-kernel learning allows for flexibility in kernel selection. This is an advantage in that kernels can easily be toggled or switched "off". Another advantage is that the loss function of multi-kernel models is convex, enabling model training using simpler, more standard optimization techniques. One of the main drawbacks of multi-kernel learning is that they are inherently slow, requiring more computational power, longer inference times, and a larger memory footprint. Ultimately, there is a tradeoff decision that must be

[1] Co-training by [23] and [123] is another technique that leverages multiple "views" of the same data to construct multiple models and will be discussed in a later section

made between a single fast and small model that may be difficult to train, or a slower large model that is more dynamic and agile to adverse data conditions [204]. This is analogous to the *exploration versus exploitation tradeoff* discussed in Section 10.

### 7.4 Modality Encoders in Unified Models

Popularized more recently is the use of separate encoders for each modality and then projecting the output of each encoder into a common embeddings space for fine-tuning of a larger model. This can be very succinctly viewed in the methods of [117] with their Large Language and Vision Assistant (LLaVA). LLaVA is a vision-language multi-modal model that provides descriptions of images based on queries about those images. In LLaVA, an image encoder (a vision transformer of similar architecture to the Contrastive Image Pre-training, or CLIP, model from [153]) extracts visual features from an image. A decoder language model (multi-layer perceptron) generates text from this image encoder. However, since the embeddings of the image encoder are not of the same shape as text embeddings used by the decoder, one must project the dimensionality of image features extracted by the image encoder to match what's observed in the text embedding space. As a result, projected image features become visual tokens for the language decoder.

Analogously, one can imagine how olfaction data and image data could be projected together to a common embeddings space to provide contextual mapping to assess which objects may be emitting certain compounds. The image-language encoder CLIP [153] is commonly used as default starting point for many vision-language models. Instead of encoding vision and language together, one can use a similar architecture to create an olfaction-vision embeddings space. This was in fact done with work from [160], a mobile application that contains an embedded olfaction-vision-language model that, as an extension to LLaVA, encodes olfaction data into a separate encoder for projection into the image-language embeddings.

### 7.5 Mixture of Experts Models

Mixture of Experts (MoE) models show promise in multi-modal learning. MoE models differ from dense models in that not all neurons are activated upon either training or inference. In a MoE, several models are assembled together with each model being trained for a specific task or domain. Each expert can be a separate task (e.g. code reasoning and mathematical reasoning) or a separate modality (e.g. vision and olfaction). A gating network handles which experts are activated for inference based on the data that is input into the model. MoE models have the added advantage in that a fault with one expert model does not require a re-training of the full MoE. One can simply edit, replace, or entirely delete the faulty model with minimal restructuring of the entire MoE.

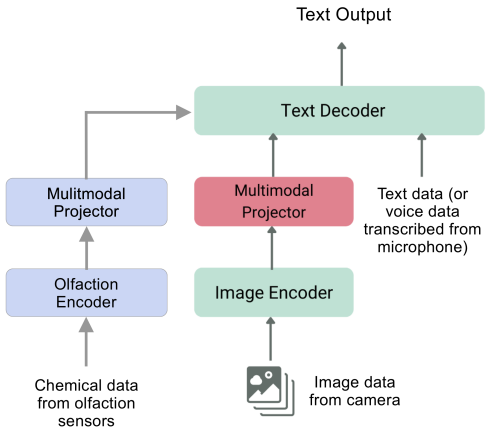


**Figure 6.** Olfaction-vision-language model (OVLM) architecture from [70].

A notable example of a MoE in broader literature is DeepSeek's R1 reasoning model [53]. In their work, they use a process called Group Relative Policy Optimization (GRPO) to generate several synthetic training examples from a few known golden examples to help optimize the MoE model. This significantly reduces the number of training examples needed to be acquired from real data by inferring them from a much smaller dataset, a technique that can be extrapolated to olfactory data.

Google DeepMind's Robot Transformer 2 (RT-2) [29] demonstrates a generalist vision-language-action (VLA) model for robots that, similar to most LLMs, was trained on internet-scale data. Research from [20] and show how these VLAs can be further optimized to run at the edge with the help of Efficient Action Tokenization [149], a process that uses the Discrete Cosine Transform to compress continuous action sequences into discrete tokens. These models have been made available on HuggingFace for use in edge robotics. Researchers from DeepMind, the Toyota Research Institute, Stanford, MIT, and others have contributed to OpenVLA, an open-sourced VLA designed to be easily finetuned on custom datasets [104], but initially trained on the Open X-Embodiment dataset [45].

More specific to olfaction, work from [70] demonstrates how multiple vision experts and olfaction experts can be combined together in a MoE to output an action for the purposes of olfactory navigation.

MoE models show promise for olfaction applications because they require, on average, four times less compute than dense models and each expert can be specifically tuned without breaking interdependencies within the larger model [8]. In addition, all else held constant, MoE models can benefit from much smaller all-up model size in comparison to their dense counterparts on similar benchmarks; this makes them more attractive to robotics applications where models need to run at the edge.

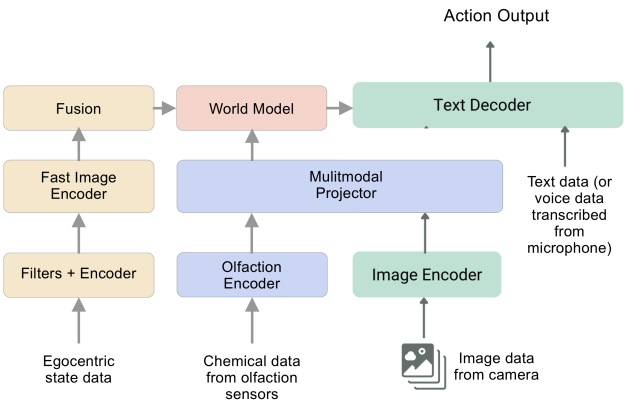


**Figure 7.** Olfaction-vision-language-action model (OVLAM) architecture from [70].

# 8 Adaptive Learning

Any statistical model is only as good as the data it is trained on. For machine learning models, the metrics achieved upon training and testing can be invalidated over time if (a) the data used to train the model was not sufficiently reflective of the entire population of data points, or (b) the population of data points changes over time. For either case, this scenario is called "data drift" or "distribution shift".

In classical machine learning, models are trained and locked upon deployment, never again allowed to change. However, the metrics achieved during training can only be assumed to hold true throughout the deployment lifetime of the model if neither of the aforementioned conditions are violated. In many cases, both of the aforementioned conditions *are* violated and the model begins to give improper results, but the user of the model may not know that distribution shift has occurred. One can imagine the critical implications of this in highly sensitive applications, such as those in healthcare and defense.

These reasons underscore the importance of adaptive learning. Adaptive learning (sometimes called lifelong learning or continuous learning) is a fairly young branch of machine learning research where agents continuously learn by encountering new tasks, gaining new knowledge without forgetting previous ones. This contrasts with the traditional train-then-deploy protocol for machine learning models, which cannot incrementally learn without experiencing catastrophic interference between consecutive tasks. Adaptive learning allows the model to adjust itself over time to accommodate evolving data inputs. In reinforcement learning, this could mean that the existing policy is allowed to adjust itself according to changing environment conditions, degrading sensors, changing behaviors of other agents within the environment, or adopting entirely new capabilities. The work of Ge, et al. in [79] lays out a concise protocol for which adaptive learning can occur.

## 8.1 Continuous & Federated Learning

It is trivial to surmise a bunch of data, feed it into an extremely large transformer, and have it recognize some sort of patterns within the data. In fact, Halevy, et al in [86] even demonstrate that said data does not even need to be sufficiently clean insomuch as it is plentiful. There is an abundance of machine learning research that has occurred over the last few decades working on solving big data with bigger models. However, in some instances, acquiring an abundance of data is not always possible. For example, in medical applications, the available data for rare diseases does not satisfy conditions for data-hungry classical machine learning. This is where continuous learning becomes prudent. Continuous learning allows the learned distribution to shift and edge closer to the true population distribution as more observations, experiences, and datapoints are acquired. [79] propose an architecture for their adaptive learning agent that comprises a larger pre-trained common "backbone" model and smaller fine-tuned modules that are adapted perform specific sub-tasks. This backbone enables agents to learn from a common representation and therefore enable faster generalization, while the smaller task-specific modules allow the agents to quickly learn new tasks and/or recognize new features. [79] label this framework as SKILL: shared knowledge in lifelong learning.

SKILL prevents the need for a large monolithic model to learn all skills together. The common backbone keeps the SKILL model lightweight and computationally efficient. The use of adaptive learning is also a factor in maintaining low resource use of the model, because the model needs less pretraining, less storage due to a smaller model size, and fewer dependencies to heavy submodules.

One can imagine how allowing a machine learning model to adjust itself can have severe implications if the adjustments are not controlled. How does the model know it adjusted itself correctly? What conditions should be put in place that govern when the model should change? We can find the answers to these questions with *federated learning*. Federated learning is a method that allows multiple decentralized models to learn on their own and be periodically synchronized with a central master model that normalizes all of the training data and observations together. The term "federated learning" was first proposed by McMahan, et al. in [126] in their 2017 work with Google.

From a swarm intelligence perspective, federated learning is a huge asset because it allows each member of the swarm to contribute to an optimal hivemind that, in turn, learns from each of the constituent swarm members. Federated learning helps limit the number of times that each constituent swarm member needs to communicate back to the master model, appropriately throttling computational resources needed for communication transactions, especially when communication is difficult or the environment denies

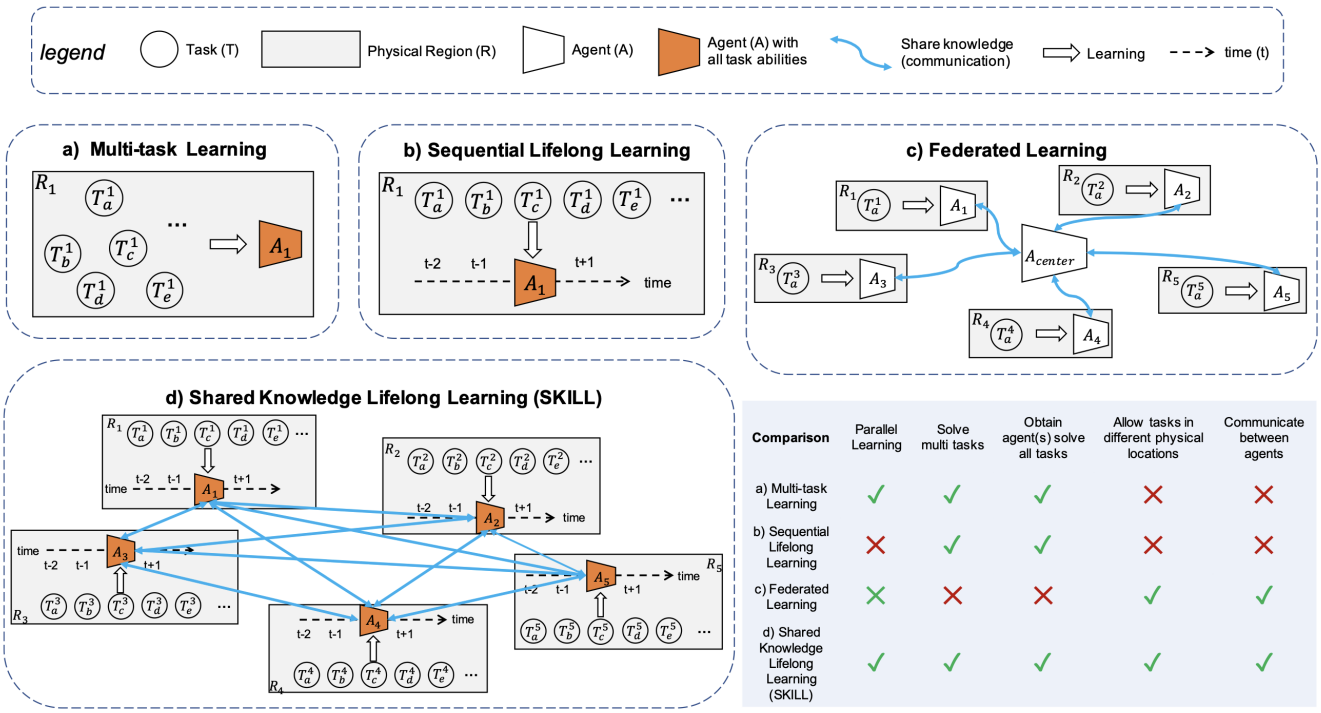


**Figure 8.** Comparison of SKILL lifelong learning with related learning paradigms. a) Multi-task learning: A single agent simultaneously learns multiple tasks in the same physical location. b) Sequential Lifelong Learning (S-LL): A single agent sequentially learns multiple tasks in one location, using specific mechanisms to prevent task interference. c) Federated learning: Multiple agents independently learn the same task in different physical locations and then share their knowledge (parameters) with a central agent. d) SKILL lifelong learning: Different S-LL agents in various physical regions each learn different tasks, and the learned knowledge is shared among all agents, enabling them to eventually solve all tasks. Bottom-right table: Strengths and weaknesses of each approach. Figure and caption adapted from [79] under the Creative Commons License.

communication among the swarm (e.g. due to signal loss). Federated learning is also a privacy-preserving feature in that it allows fewer data transactions to occur and more isolated edge models limiting the opportunities for adversarial attacks. Secondly, federated learning allows each agent of the network to learn a slightly different part of the distribution. The synchronization back to the master model allows all other edge agents to leverage the experiences of the other agents without directly having to have those experiences themselves. This allows for "multi-threaded" or "distributed" learning in that each agent is learning different details about a common task.

From an adaptive learning standpoint, this is extremely attractive as it allows agents to adapt to their environments in a controlled manner. For example, through the lifelong learning framework in [79], agents can quickly learn new skills through federated learning by maintaining a common backbone model and several finetuned models on each skill. Each agent has a separate finetuned model for specific tasks, but the backbone is a hivemind model contributed to by all other agents in the swarm. Figure 8 illustrates how the SKILL lifelong learning framework from [79] uses federated learning to keep machine learning models light and adaptive.

Federated learning not only helps spread information among constituent members of the swarm. It also promotes controlled and stable learning among distribution shift. As [170] discuss in their work with prototypical networks, gradient explosion, catastrophic forgetting, and other means of "unlearning" can occur when uncontrolled active or transfer learning occurs. Federated learning allows the swarm of agents to, indirectly, shift the prototypical network to align with the true network by taking the weighted average of the prototypes learned by each agent. This helps ensure that any shift in network modeling is generally moving in the direction of the true data distribution.

## 8.2 Self-Adaptive Networks

The previous section discussed a style of learning that allows the weights of a network to continue to change throughout its use. Another approach to continuous and adaptive learning is through the use of columnar constructive networks (CCNs). Originally established by Javed, et al. in [97], CCNs

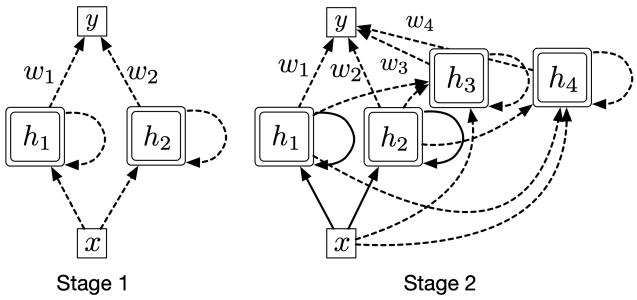


**Figure 9.** Columnar-Constructive networks (CCNs) combine the ideas from Columnar and Constructive networks. Old knowledge is preserved in the frozen layers and new knowledge is acquired in new layers (columns). This allows the network architecture to adapt to the task or pattern complexity. Figure adapted from [97] under the Creative Commons License.

begin training with a shallow neural network, typically only a couple of layers. The error of the network is monitored and, when enough evidence has amounted that the network can no longer improve prediction quality, another layer is added. Each layer is considered a "column". As the next layer is added, all preceding hidden layers are frozen, effectively preserving knowledge over the data learned thus far. As learning continues, only the weights of the recently added layer are adjusted. This process continues until some threshold of error is accomplished consistently, and the addition of layers may be throttled by a temperature parameter. Javed, et al. mention the optimization of the temporal difference (TD) error and emphasize the use of the technique in reinforcement learning. In contrast to simply adjusting weights of a constant-size network, the use of CCNs enables a network that grows to the complexity of the task or pattern.

We find the CCN technique particularly compelling for machine olfaction. As a robot navigates, the scent it is tracking may evolve or change as its concentration ebbs and flows and it reacts with other compounds in the air. Maintaining snapshots of the plume in the frozen layers could capture a temporal snapshot of a scent as the robot pursues it to the target source. More on plume tracking is discussed in Section 6.

### 8.3 Task vs Style vs Modality Transfer Learning

One goal of adaptive learning is to develop AI systems that can continuously learn to address new tasks from new data, while preserving knowledge learned from previously learned tasks. Discussed above, this is known as multi-task learning or task transfer learning. Also previously discussed is multi-modal learning where multiple data sources are jointly learned together into a single reward signal. There is another type of knowledge transfer in adaptive learning called *style transfer learning*.

Style transfer learning can be generally pictured with the following anecdote: Two agents have an identical goal of navigating a course to the finish line. Both agents know the route to the finish line with 80% confidence but one agent has an exploratory style and the other agent has an exploitation style. In other words, one agent is going to complete the course as fast as possible and have full confidence that it's policy is correct, deviating as little from its policy as necessary. The other is going to attribute less confidence to its policy and explore alternative routes to the finish line in the event that one route may lead to the finish line more quickly. Both agents have identical sensors, goals, and reward signals, but their styles of achieving the desired goal is slightly different.

The work of [138] illustrates this concept well. In their manuscript, they propose a framework called *MultiCriticAL* that shows how actor-critic reinforcement learning methods can be extended to support different styles via the use of multiple critics among a single actor. Actor critic methods are temporal different reinforcement learners that represent the policy independent of the value function through the use of separate policy and memory mechanisms. The policy mechanism is known as the actor and is the one selecting actions. The critic mechanism is the estimated value function because it "criticizes" the decisions made by the actor. The back-and-forth exchange of state-action trajectories between these mechanisms facilitates learning in the form of temporal differences between the actor's decisions and the critic's scrutiny. Learning always occurs on policy. A diagram showing the MultiCriticAL framework from [138] can be seen in Figure 10.

In typical actor-critic learning, the value functions are learned by a single network with a single output node for all tasks. One problem with using a single learned value-function for multi-task learning is that one value-function representing multiple values may enforce continuity between the learned values. If this occurs, it could compromise the quality of the learned values and resulting policies, decreasing the quality of the learned values and the consequential performance thereof. [138] suggest separating the learned values per style, or task. The continuity between task values can be dismissed, and the values for each task specifically learned without relying on the critic's policy to learn the distinctions between each.

The MultiCriticAL framework shows how multi-valued critics, and therefore multi-task reinforcement learners, may be effectively represented by separate independent critics, or a single critic network with multiple heads. In the latter scenario, each each head details a different learned task-value. This allows for some learned representations to be shared among the value functions if the tasks to be learned are somewhat similar and using separate critics ends up being too computationally expensive. This can be more explicitly seen by the diagram in Figure 11.

The claims by the authors of the MultiCriticAL framework are validated by their results. Their study analyzes the difference in path following for single-style soft actor-critic (SAC), multi-task SAC, and both the multi-network and multi-head flavors of MultiCriticAL. [138] claims to achieve a 56% increase in multi-task performance with MultiCriticAL in comparison to the predicate methods, all while using smaller neural networks to construct their policies.

While Mysore, et al. proposed an actor-critic architecture with many critics in [138], Li, et al. in [114] propose the converse: an actor-critic architecture with multiple actors for multi-task learning. While their approaches are similar in exploitation, there are some implementation differences. Intriguingly, [114] do not specifically emphasize their methodology for multi-task learning, but rather for the use of more accurately learning single complex tasks by breaking them down into smaller sub-tasks. They advocate the use of their model for the purpose of weighting the different Q-values of different actors for the same task performed, and then using these values to more accurately construct the single-task policy. The authors illustrate their multi-actor method with deep deterministic policy gradient (DDPG) and twin delayed DDPG (TD3) architectures. What they call the *multi-actor mechanism* or *MAM*, they demonstrate how it can hedge against estimation bias, enhance exploration, and provide state-of-the-art results in several of tasks from the MuJoCo battery [28].

[114] demonstrate the use of multiple actors among a shared critic with the intent for each critic to learn different sub-distributions of the overall task distribution. However, the work of Zhang, et al. in [202] extends this work to more accurately reflect the converse of the MultiCriticAL framework from [138]. They demonstrate how the critic can act as a knowledge transfer mechanism among many independent actors, where each actor learns a different specific task. One note that the authors emphasize is that the hyperparameters and architectures of each actor model are not necessarily identical. This allows the shared-critic architecture to learn largely heterogeneous tasks through the use of differently architected actors.

### 8.4 Multiple Methods, One Objective

Each of the aforementioned methods contribute a different angle to multi-task learning. Whether through the use of many critics as in [138], many actors as in [202], or other methods such as those proposed by the federated learning work of [126], multi-task learning is proposed as being an effort that is easiest accomplished via multiple models. This segues nicely into multi-agent learning. If multiple models can be built to enable a single mechanism to learn multiple tasks, each of these models can also be analogously leveraged to enable multi-agent cooperation. In this manner, multi-task methods can be easily extrapolated as multi-agent and swarm learning methods. For more heterogeneous swarms, the work of [138] and [202] can be used, while the multi-actor model by [114] can be leveraged for more homogeneous swarms. There is a host of shared knowledge between multi-task and multi-agent learning that can be used for swarm design.

### 8.5 Synthetic Data

In the absence of plentiful data for model training, generating synthetic data from the data that is available is a reasonable option to scope a sufficiently large dataset. Gaussian perturbation methods from [156] are an option for generating synthetic data from seed data, with the acceptance that having Gaussian modes around each true training sample is an acceptable approximation of the real data distribution. GRPO from [53] also shows a lot of promise in optimizing reward models.

When generating synthetic data, one must be cognizant of model collapse. Too much synthetic data generated from previously trained models (versus synthetic data generated from ground truth data or golden queries) can lead to vanishing gradients and model collapse if not controlled. Methodology from Seddik, et al. in [163] denote several methods to hedge against this. They delineate multiple recipes and limits for mixing both synthetic and real data together during training to prevent model collapse.

In summary, synthetic data generation is a very credible technique to use for olfactory data, but it must be done so with control and caution.

### 8.6 Plasticity

Converse to model collapse is plasticity loss. When a model loses the ability to adjust and train (its plasticity), the amount of data and its diversity make negligible impact on the model, regardless of the magnitude. A very detailed survey from [58] illustrates conditions under which plasticity in deep learning models begin to occur. Loss of plasticity is of primary concern with continuous learning. When a model is able to constantly update itself, gradients become smaller and smaller until no more updates can be made, and plasticity itself converges to zero. One remedy for this is to implement learning rate decay, such that plasticity is much larger in early learning than it is in later learning. Both in and outside of machine olfaction, robots will require the ability to continuously learn from few examples and it is paramount that technical architects consider strategies for maintaining plasticity during adaptive learning.

### 8.7 E-Prop and Eligibility Traces

One of the most computationally expensive procedures of machine learning algorithm is the weight update step. Typically, this is conducted through the common backpropagation algorithm proposed by Yann LeCun in 1988 [109]. It has since become a staple of machine learning. However, backpropagation requires the computation of expensive first-order derivatives in order to adjust weights within a neural

network. To assess whether convergence has occurred, many ML training procedures compute second-order derivatives of the Hessian matrix, which is even more computationally expensive.

An alternative weight update algorithm to backpropagation called *E-prop* was developed by Bellec, et al. in 2019. E-prop performs weight updates via approximation and through the use of eligibility traces.

The synaptic update rule in E-prop can be illustrated as follows:

$$\Delta w_{ij} \propto \sum_{i} e_{ij}(t) \cdot \delta_j(t) \tag{3}$$

Where $e_{ij}(t)$ denotes the eligibility trace for the synapse between neurons $i$ and $j$, $\delta_j(t)$ is the learning signal that provides feedback, and $\Delta w_{ij}$ represents the weight update between neurons $i$ and $j$.

Initially proposed for the use of spiking neural networks in neuromorphic computing, its much lower computational demand makes it an attractive optimization algorithm for adaptive learning. Eligibility traces are conventionally used in temporal-difference reinforcement learning algorithms to assign credit to the most probable actions based on recent exploration history. This concept can be extended beyond reinforcement learning to enable adaptive learning models to "forget" irrelevant patterns and assign higher credit to more recent observations.

Backpropagation is an exact calculation of the weight updates and E-prop is an approximation of the weight updates which dissuades its use among the community. However, exactness in adaptive learning can actually be a hindrance to performance as the model is not mature enough to provide exact assessments. As we will show in Section 10.7, approximation through expectation can be more beneficial than exactness in learning under extreme uncertainty. Analogously, work from van Hasselt and others from DeepMind [183] show how eligibility traces can be adjusted to be more informative of expected outcomes than singular exact outcomes when informing the output of a machine learning model. We suspect these "expected eligibility traces" to be most informative for adaptive learning in machine olfaction.

Adaptive learning also demands computational efficiency in order to make live updates. This is especially true for machine olfaction applications since most olfactory sensors run at the edge. While little comparative research exists comparing e-prop to backprop through rigorous ablation studies, even less research exists addressing the use of e-prop in machine olfaction. The benefits that e-prop and expected eligibility traces provide for adaptive learning are expected to significantly benefit the highly dynamic machine olfaction tasks such as plume tracking.

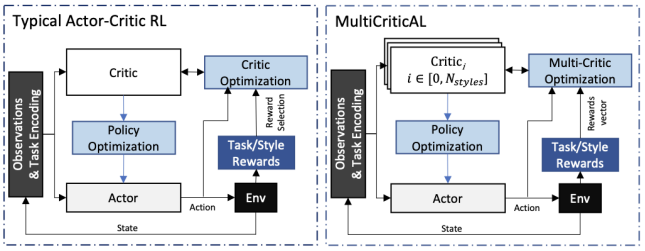


**Figure 10.** A diagram of the *MultiCriticAL* architecture. MultiCriticAL diverges from the common use of a single critic and instead uses multiple critics, a separate one for each task and/or style learned. The authors illustrate MultiCriticAL's success in training multiple distinct behavior styles in various games, including Pong and UFC. Figure adapted from [138] and used under the Creative Commons License.

## 9 The Influence of Game Theory

Typical statistical learning methods require training on some set of prior data in order to construct a model that can be worthy of optimizing for some objective. The utility of this data in training a model implies that, at some point, the distribution (or environment) was observable to a degree that allowed this data to be acquired. In a scenario where one is not afforded the luxury of knowing the environment beforehand, many of the assumptions regarding data-greedy learning become invalidated and therefore do not contribute to the construction of a useful model. Vidya Muthukumar focuses on this premise in her dissertation Learning from an Unknown Environment, the thesis of her Doctorate of Philosophy research at the University of California at Berkeley.

In her paper [133], Muthukumar establishes reasons behind why learning from an environment with no known priors is difficult and the precautions one must take in making certain assumptions. She attempts to demonstrate principles that may constitute a framework for facilitating learning in an unknown environment that is almost as optimal as though we had known the environment beforehand. Particularly, she emphasizes the potential of such a framework for real-time adaptive learning. She denotes four behaviors that an agent can possess while learning: stochastic, adversarial, competitive, and cooperative. A thorough analysis is conducted on how to learn in the presence of the first three, while the latter behavior is acknowledged as a natural extension of her work to be evaluated elsewhere, leaving us with some motivations on how it might be pursued. Throughout the paper, Muthukumar makes consistent references to the Frequentist versus Bayesian paradigms and where the inclusion of one methodology over the other may be more adept to addressing the certain problems. Fundamentally, Muthukumar establishes a framework for optimally learning from an unobservable environment and how to develop a policy to do so in real-time while adapting to different environmental

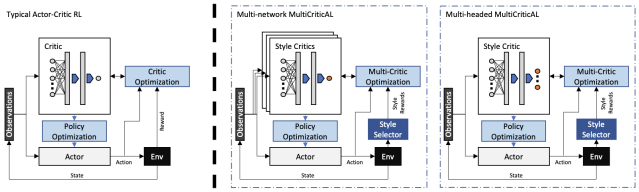


**Figure 11.** A diagram of the multihead version of *MultiCriticAL* architecture. This can be used as a more computationally efficient version of the architecture if desired. Figure adapted from [138] and used under the Creative Commons License.

conditions by exploring single-age learning in the context of multi-agent game theory.

### 9.1 Learning in an Unknown Environment

The dissertation is constructed as follows: Chapter 1 provides motivations behind the interest and relevance of adaptive learning algorithms in an unknown environment; Chapters 2 and 3 provide the foundation for selecting an adaptive model from which an agent can learn with respect to stochastic and adversarial behaviors as a function of regret; Chapter 4 shows how the framework established by its successors can be extended to competitive behavior learning; Chapter 5 illustrates how the previous chapters' framework for single-agent learning can be extended to multi-agent learning; and Chapter 6 concludes the thesis with some remarks on future work. Throughout each chapter, Muthukumar does an excellent job of successively building on her previous arguments with insightful ties to prior work in the field and how her approach can address their shortcomings.

It is worth noting that the the dissertation provides arguments with heavy respect to game theory versus statistical learning, so exhaustive evaluation of hyperparameters and their influence on performance are not discussed in great detail nor should they be. Therefore, the assessment of Learning from an Unknown Environment presented here will contain minimal dialog and diagrams around such performance evaluations in comparison to other papers focused on statistical learning.

She references Paul Milgrom's text [130] as one of the earlier motivators behind this effort. She also relates the applicability of constraint satisfaction through her mention of SAT solvers [111] [140] and market matching [99] [118]. Lightheaded Regulation, proposed by Kristen Ann Woyach [196] also indicated as another significant motivator due to its modeling of asymmetric user interaction. Only a single allusion to reinforcement learning is made with [177] which was originally quite surprising since RL has many natural extensions of this paper as a whole. However, the author's goal is to focus on the development of a game theoretic framework not weighted by classical statistical learning techniques, so the lack of these references are sensical.

The author frames many of her arguments in the context of minimizing regret. One can define regret as the difference in loss between the action just taken, and the optimal action that should have been taken in hindsight according to the optimal behavioral policy. Early concepts of regret were first posed by Jim Hannan [87] and David Blackwell [21] in the context of how the approachability of convex sets evolves over time. The "no-regret" property is a cornerstone in the author's proposed framework for learning in unknown environments and the concept seems to cross over into economics theory [88] as well, which is unsurprising given game theory's heavy presence throughout the paper. She gives further support for regret by referring readers to [39] [158] [105] that illustrate how "Follow-the-Leader" style algorithms can achieve minimal regret of O(1).

One of the main goals of the author's dissertation was to show how adaptivity could be used to minimize regret between stochastic and adversarial environments as a function of model complexity. Early players in adaptive online learning were illustrated by [94] [39] and the author quantifies this well by discussing some of the more modern work done in the space by [182] [129] [68]. The author grounds her model selection arguments with principles in structural risk minimization (SRM) originally presented by Vladimir Vapnik and Alexey Chervonenkis in their 1974 paper [184]. Muthukumar leverages work from [136] to couple the SRM framework with the AdaHedge algorithm from [39] [158] [182] as part of her model selection methodology. She extends this evaluation by showing how the coupled algorithms can select adversarial models by assessing the cumulative variance in the algorithm for a second-order regret bound.

In Chapter 3, the author discusses how her methodology in Chapter 2 constructed what was effectively a meta-framework of bandits "corralled" together—a direct extension of the work in [4]. Remarks around other ensemble techniques proposed by [108] [67] are also referenced in the context of a more optimized meta-bandit mixture of experts. Eventually, the author proposes two algorithms to evaluate her arguments largely built on the Stochastic and Adversarial Optimization (SAO) method presented in [32].

References to asymmetric learning [145] are given later on as the author explains the importance of credibility [172] [124] and commitment [161] in the context of repeated interaction. Notions of commitment (especially in the context of game theory) draw references from Stackelberg [173]. Stackelberg demonstrated the power of committing to a single strategy in a special kind of one-shot interactive game between a designated leader and a designated follower (called a Stackelberg game). There is significant discussion around Stackelberg equilibria that provokes evaluation of sub-game perfect Nash equilibria (SPNE) which, in the context of this paper, can be viewed as a series of interactions between an agent that is viewed as a "leader" over other agents "followers" [106] [131]. Arguments about the limits and benefits

of Stackelberg assumptions in repeated interactions ensue which motivate the author's final single-agent model [69] [137]. Many principles of [33] have credence for discussions that emerge thereafter, especially regarding scenario complexity as a function of learnability. However, there does not seem to be any references to this inductive-logic methodology. This discussion of single-agent learning under a Stackelberg authority teases work on a similar multi-agent strategy. [88] [69] [31] provide insight into how the monitoring of the moving averages of players' strategies can be an effective means of predicting their behavior. Correspondingly, sources [51] [3] [115] extend this research to evaluate the performance of players employing limited mixed strategies, or "last iterates." [51] and [1], along with the author in a separate work [134] give a more recent update on last-iterate convergence and divergence since the dissertation was released.

### 9.2 Frequentist vs Bayesian Statistics

The tradeoff between Frequentist and Bayesian thinking is important in all robotic learning environments. When Bayesian conditions are violated, Frequentist techniques must be employed to construct the prior distribution for which decision making can occur. [133] argues that frequentist thinking motivates many of the principles behind game theoretic methodologies. She continues this support by showing how her own methodologies in inferring unobservable distributions also support the frequentist angle. However, she makes several cases on where Bayesian approaches provide more merit over its counterpart. In Chapters 2 and 3, she explores how agent performance is influenced by (a) establishing assumptions over each paradigm, and (b) mathematically proving the limits behind the performance of each as a function of agent regret and behavior. These construct the foundation for her arguments in the following chapter where she demonstrates how her frequentist framework allows a learning agent engaged in a non-zero-sum game (a Stackelberg game) to approximate other agents' behavior by assessing their respective tradeoffs and incentives, and then influencing its own behavior as a consequence. In turn, she shows how this allows the primary agent to reach the Stackelberg equilibrium of the game.

The emphasis on the frequentist mindset is not one that is as prominently spoken about in statistical learning or robotics. [133] makes several great cases around why and where frequentist methodologies succeed Bayesian, especially in light of how brittle Bayesian models become when their assumptions are not validated. There is strong evidence to support these statements in dynamic environments—Bayesian learning requires a prior distribution to be established which implies an experience has been encountered before. For tasks and data never before seen, a robotic agent cannot leverage a prior distribution, but must construct it. In many of Muthukumar's arguments, proofs, and lemmas, she focuses on how the frequentist framework does not require the need of a common prior and how this promotes convergence of many of the proposed algorithms. However, one concern with these arguments (and this concern is not pointed directly toward Muthukumar - it seems that many frequentist arguments do not define this well) is that the start of agent learning is not addressed. If one assumes a frequentist model, one has to construct this model to immediately begin making decisions. Where does this model come from? It must come from observations of the previous game states. The initial few turns of an agent inferring from a frequentist perspective will be highly volatile. Granted, this volatility is a function of the state space and action space sizes, but not until a representative sample of moves is acquired will the agent begin constructing an optimal model that out-competes a Bayesian one. Now, the counter-argument to this is that, if one is not reasoning about an observable environment, a reliable Bayesian prior cannot be constructed because there is no knowledge to construct it with. This is a fair case, but it is worth pausing for a moment to think about whether it is it more destructive to employ a "guessing" frequentist model or establishing a (potentially) invalid common prior to facilitate the first few game moves. The answer to this may depend on the problem, the bias in an agent's sensors and the current agent state.

Additionally, the Muthukumar indicates in multiple places that the maximum likelihood estimator (MLE) is used to predict agent behavior under the frequentist methodology. However, this assumes that the agent will continue playing with "expected" behavior. What if the secondary agent is intentionally deceiving the primary agent with stochastic moves in a "Follow-the-Leader" behavior, only to change to adversarial behavior as the follower realizes the leader is close to the goal? At this point, the MLE may not prove optimal. In Section 2.2, Muthukumar briefly discusses how adversarial behavior could be interpreted as stochastic behavior by analyzing an example using the binary sequence prediction problem. This example excellently sets up her argument for the need for model adaptability in an unobservable environment. However, it seems like some sort of dialog around the advantage of a Bayesian prior would be helpful here to make the agent continuously self reflect, "What is the probability I am assessing the behavior of my fellow players correctly given what I have observed in their behavior thus far and given that my policy is 100% correct?". At this point in my analysis, I did not think that there was compelling evidence presented to suggest that the frequentist paradigm appropriately rectified a change in other players' behaviors, especially since this partially grounds the rest of the author's framework. Regardless, the author does recognize that this could be fatal to a learner, and moves on to show how model selection and re-selection should occur in the presence of this.

Almost immediately after this, the author shows how not only is model adaptability needed, but data adaptability is needed as well. The investigation into this is admirable and quite prudent, and could have been easily overlooked. This highlights a key point in adaptive learning, and the arguments are even more intriguing in the context of a frequentist approach. Muthukumar's model selection processes thereon illustrates the best selection criterion with respect to both behavior and data which, ignoring the above concern, was otherwise well defined.

The tradeoff between frequentist and Bayesian thinking is a crucial one for adaptive learning because learning among no data will deprecate the utility of data-hungry machine learning methods. Since the field of artificial olfaction lacks large high-quality training datasets, we observe that it is important to understand how to use frequentist methods to construct an initial model for olfactory tasks that can become the prior for more sophisticated techniques upon gathering adequate data.

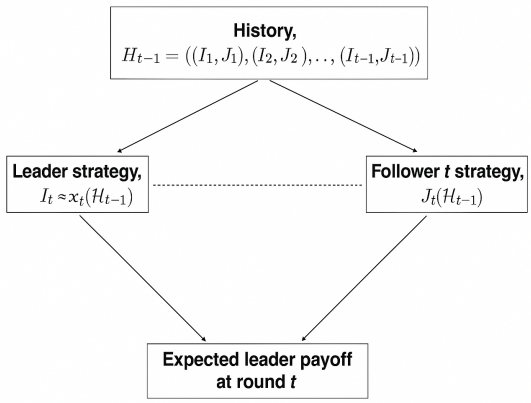


**Figure 12.** Illustration of the a Stackelberg game at round $t$ between leader and follower $t$, both of whom observe history of play $H_{t-1}$. The dotted line indicates that leader and follower $t$ play simultaneously.

### 9.3 How Behavior & Regret Influence Adaptability

Throughout the text, [133] builds much of her framework on the concept of minimizing regret. While this is not a new term, she postulates it in some in some new ways. For example, she pairs regret with model complexity in order to select the least complex model as a function of behavior. She alludes to the paper of Vapnik and Chervonenkis [184] in defining regret as general principle calculated by the difference between the action an agent took (or will take) and the action it should take under the optimal policy. She makes another great case for providing data adaptability by specifying a scenario where the environment may be mis-specified. She defines a mis-specified environment as one that is truly stochastic, but not observed as such by the learning agent. The point is not explicitly made, but one can imagine that an agent may interpret the environment as stochastic due to its lack of observability, but it is in fact adversarial and the agent is being deceived, perhaps intentionally. This scenario would obviously constitute a large amount of regret and the author indicates that one can bound the problem through second-order-bounding, a process also pursued by [14][15][18].

Muthukumar posits two bandit models in Chapter 3 - one simple model and one complex model. She compares the tradeoff in regret between both algorithms over different model selection methods: OSOM (Optimistic Selection of Models), OFUL (Optimism in the Face of Uncertainty Learning) [12], and UCB (University of California at Berkeley) [132]. Ultimately, the author rules that the simple model, a mixture of K-arm bandits, can approximate the complex model, a contextual bandit model with sufficient accuracy while also sufficiently bounding the the regret[2].

The potential confusion between stochastic and adversarial environments is worthy of focus for a moment. In Chapter 1, the author specifies that the detection of adversarial and stochastic behaviors favor frequentist approaches, but that the detection of cooperative and competitive behaviors favor Bayesian methods. This augments the concerns in section 3.1 above, where a pure frequentist approach may be insufficient for capturing the information needed to produce a minimum-regret model. She makes a statement on page 141 that she assumes all other agents are using frequentist principles for their inference. This seems dangerous and causes some suspicion of fragility in the framework she is proposing, as this assumption cannot be guaranteed in an unknown environment. She provides a thorough comparison of Bayesian and frequentist methodologies for her final single-agent model, but it is still unclear if the model would minimize regret under the scenario where other agents are not exuding frequentist strategies. Regardless, her framework is helpful from an adaptive learning perspective because adaptive learning inevitably takes strong influence from frequentist principles.

### 9.4 Last-Iterate Divergence

The author discusses Stackelberg games [26], equilibria [30], and commitments [27] at great length. The modeling of a Stackelberg game frames the final model that the author proposes for a single agent learner in an unknown environment. A Stackelberg game models a one-shot interaction between two agents: a leader and a follower. In this context, the leader is a data generator and the follower is a learner. The follower can learn the behavior (the commitment) of the leader from simply observing the leader's actions. The effectively reveals its strategy to the learner in this manner. Typically, Stackelberg games require that the leader commits

[2]The proofs for the bounds of regret on both simple and complex models can be found on pages 116 and 118 of the dissertation, respectively.

to a pure strategy, but the author postulates that a mixed strategy can be employed here. Figure 12 shows a diagram of this leader - follower interaction. Muthukumar argues that this model of commitment paired with regret minimization significantly benefits the leader because the leader possesses perfect knowledge of its own utility function; it uses this power to maximize its expected reward through exploitation of information asymmetry.

At some point, the leader's commitment to a certain behavior will build "credibility" among the follower, or a reputation of acting in a specific way. The follower will then begin making decisions at time t based on the history acquired at time , modeling its own policy after . The leader is then incentivized to deceive the follower's learning process once this credibility is realized and begin deviating from its historical strategy. It is thus in the leader's best interest to incorporate some randomization into their behavior to build more robust credibility. Put differently, the leader's slight deviation to a higher regret behavior (through deception) may actually manifest the true no-regret behavior if the follower continues to be deceived because the leader may coerce the follower into a non-optimal strategy.

Here, the author completes her discussion on single-agent learning. Her arguments are well-structured and she provides significant mathematical proof. However, one of her concluding comments is that the above framework will hold true for "novice" and "unintelligent" followers. From one level of analysis, it should be defined what constitutes a "novice" follower. In this sense, one agent may "follow" another agent in its actions under certain scenarios, especially if the latter agent truly does have perfect information and the former is largely exploratory. However, the conditions under which these events occur do not seem generalizable to the degree that the author promotes. Furthermore, the author argues that Bayesian methodology would hold minimum value here because a sufficient number of data samples will "drown out" the effect of a prior (page 138). Is not the methodology of Bayesian learning to consistently update the prior with a posterior assessment that reflects new knowledge? The footnote on page 138 slightly relaxes her stance in this regard where she suggests that there is strong evidence to support that Bayesian and frequentist methodologies complement each other in problems such as this and multi-arm bandits.

The rest of the text is devoted to extrapolating this learner to "two-sided" learning in unknown environments. While Muthukumar concludes the previous section showing her final model of a no-regret one-sided learner in an unknown environment, she shows how multiple learners employing this same strategy together provide divergent results. We can consider the scenario of a single learner above exuding last-iterate (time-average) convergence to its solution. Through an extensive series of proofs, the author shows that the no-regret and last-iterates properties show strong adversarial tendencies toward one another such that no-regret strategies actually imply last-iterate divergence in two-sided learning [3] (page 235). In both the deterministic and stochastic cases, one can observe that the multiplicative weights exhibit highly volatile behavior, even with the optimal no-regret rate [4]. The author cites stochasticity inherent in player's realizations as a critical reason behind this divergence [3]. She notes:

*"The primary distinction in our work (as well as the settings of An et al., Shieh et al. and Blum et al.), is that the manifestation of the uncertainty is itself random. Thus, a unique component of our results involves directly reasoning about the stochasticity of the follower response."*

Other work proposed by [118] has been recently established showing last-iterate convergence with no-regret learning under certain constraints [40]. In my review of Lei's work, it does not seem immediately apparent that what his paper postulates conflicts with what Muthukumar's suggestions above about the impossibility of the convergence. Both apply no-regret learning, but the former documents a specific case over Min-Max optimization while the latter's is tailored specifically to mixed strategies.

### 9.5 Evidential Uncertainty

With very small datasets, it can be easy to overfit any statistical model as the sample distribution is at a large risk of not reflecting the entire population. We have a strong interest in hedging against overfitting the models proposed above to show the honesty of our results. Given the nature of our application, acquiring thousands or millions of data points is not yet practical due to the stage of our product development, funding, and scientific validation. To alleviate this risk of overfitting, we go beyond simply training a machine learning model by also quantifying the model's uncertainty using a concept called evidential deep learning. In machine learning, models are trained to learn the average correct answer for a given input, but they do not model any noise or uncertainty in the input whilst doing so. There are two types of uncertainty that can be modeled: aleatoric and epistemic uncertainty. Aleatoric uncertainty refers to inherent randomness present in the data that can be the result of sensor variations (drift, heating, saturation, etc.).

Epistemic uncertainty addresses uncertainty in missing data or, more precisely, how well the distribution of the sample population and training data probabilistically relates to the distribution over the entire population which is typically unknown or difficult to measure, hence the need for a statistical model. Evidential deep learning (EDL), or evidential uncertainty modeling, is a method for modeling epistemic uncertainty within a machine learning model through the

[3] In a publication that occurred after her publishing of this dissertation, the author claims to show that it is impossible to employ a mixed strategy with no-regret learning and achieve convergence. See reference [135]

[4] Regret rates are defined on a continuous scale between 0 and 1, where 0 indicates full regret and 1 indicates no regret (e.g. we execute the optimal move according to the optimal policy every move).

construction of belief masses—probability vectors representing evidence of data supporting any one class. These belief masses are different than classification probabilities or output logits typically seen from softmax or sigmoid layers in regular machine learning models in that belief masses give the model the ability to effectively say, "I don't know". EDL extrapolates from Sutton's method of learning to predict from temporal differences [176] and is heavily inspired by Dempster-Schafer Theory of Evidence (DST), a hypothesis for generalizing the Bayesian theory of subjective probabilities. The Dirichlet distribution—the probability density function for the prior of the multinomial distribution—models belief masses for a multi-class problem; in this case, we are attempting to predict two classes and the binomial-variant of the Dirichlet distribution is simply the Beta distribution.

Under the Subjective Logic Framework [164], the assignment of belief masses infer that the belief of the truth can be on any of the possible given states or classes, giving an overall uncertainty mass $u$ over $K$ possible states that sum to 1 in accordance to the following:

$$u + \sum_{k=1}^{K} b_k = 1 \tag{4}$$

Where the uncertainty mass $u$ is a positive definite value, $K$ is any number of states greater than 1, $e_k$ is a positive definite value indicating the evidence derived from the $k^{th}$ class, and bk represents the belief of class K according to the following:

$$b_k = \frac{e_k}{S} \tag{5}$$

The uncertainty mass $u$ is quantified as the number of classes over the sum of the aggregate evidence $S$, defined below:

$$S = \sum_{k=1}^{K} (e_i + 1) \tag{6}$$

In short, for both classification and regression, evidential learning simultaneously assesses uncertainty about both the data and the model. EDL learns to approximate aleatoric uncertainty from the data and epistemic uncertainty from the model. For any input, the network is trained to predict the parameters of an evidential distribution of which models a higher-order probability distribution over the individual likelihood parameters associated with both the data and model [7]. This concept is more concisely represented in Figure 13. EDL applies to adaptive learning and machine olfaction in that, given the low quantity of data available for training olfactory models, EDL methods can be applied over smaller datasets to help control the amount of bias learned by a model.

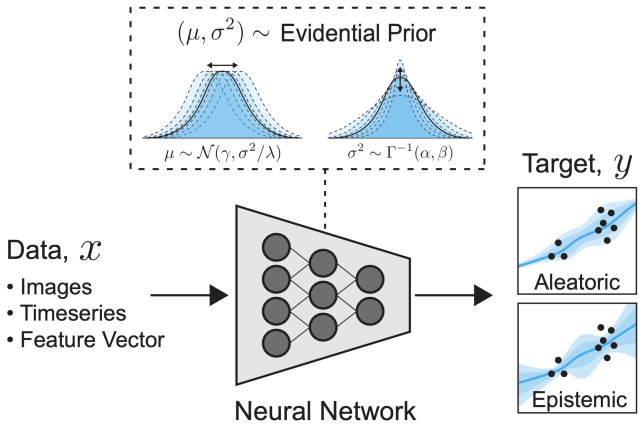


**Figure 13.** Evidential learning simultaneously learns a continuous target along with aleatoric data and epistemic uncertainty from the model. Figure and caption adapted from [7] and used under the Creative Commons License.

### 9.6 Application to Adaptive Learning

In general, concepts in game theory share collinear goals with concepts in reinforcement learning, and this paper is no exception. A natural extension of the author's paper would be to align her methods with a simple Q-learner and assess their collaboration in navigating unknown environments. Additionally, in the conclusion of section 3.3 above, the author showed that the stochasticity of follower response played a critical role in the last-rate divergence. So an immediately obvious appendage to the work here would be to see if there is a way one can build an auxiliary model to approximate [5] the noise of the response of the follower in order to discourage divergence from occurring as quickly.

Finally, the paper's concepts do seem to allow for some organic cooperation with inductive logic programming. In many areas, the author alludes to a set of rules that govern learning behavior. She regularly begins the evaluation of her models with simple games that could show strong affinity for the techniques proposed by Copper et al. with their framework for the learning of simple games [47]. Muthukumar's specific contributions toward learning adversarial behavior and inferring the unobservable could hedge against the predicate explosion concerns proposed by Minton [132] that exist among today's rule-based game playing.

## 10 Swarm Intelligence

A swarm is a group of agents ideally working toward a collective goal - a multi-agent network. The work of Craig

[5]We note "approximate" because Muthukumar already suggests in her 2022 publication that it is allegedly impossible to fully characterize this response while employing no-regret learning and mixed strategy behavior. So perhaps an approximation can occur, and, hypothetically, perhaps there are methods that can be developed to discourage divergence as quickly such that a learner has time to deploy a new strategy.

Reynolds in 1987 [157] with his Boids algorithm showed how complex behavior could arise within groups of simple particles obeying only a handful of simple rules. His work extended, in a way, the cellular automata concept proposed in Conway's Game of Life in 1970 [77]. These works incentivized more research into swarm intelligence and the computer simulation of social interactions among artificial agents. Muthukumar's work in [133] highlights four distinct behaviors that appear in multi-agent game theory: competitive, cooperative, adversarial, and stochastic. Throughout the survey, we leverage her taxonomy by grouping our observed behaviors into the same four classes and provide evidence to support many of her game theoretic points but in the context of reinforcement learning and active sensing. [200] emphasize the collaboration that groups of proximal policy optimization (PPO) agents foster in multi-agent games.

The research of [169], [75], and [93] demonstrate many principles of adaptive learning through the tracking of odor plumes, an extremely dynamic and high degree-of-freedom (DoF) task. They illustrate how RL agents learn to maintain a statistical estimate of a scent trail trajectory analogous to the method in which rodents and other terrestrial animals track scent trails. They also establish some theoretical limits on how quickly these scents can be tracked due to fundamental geometric constraints.

### 10.1 Reinforcement Learning

Reinforcement learning is a single-agent reward-based machine learning model for control problems. A reward signal is defined for a specific task and an agent learns to maximize this reward signal as it explores its environment and exploits actions. Careful consideration is given to the design of this reward signal to ensure the agent learns the appropriate behaviors. Imitation learning and inverse reinforcement learning fall under the broader category of reinforcement learning. In their paper, Abbeel and Ng describe inverse reinforcement learning (IRL) as the task of determining a reward function based on the observed behavior of an 'expert' [2]. On the other hand, Huang et al. define imitation learning (IL) as the process of deriving a policy from expert behavior demonstrations [89]. Although both IL and IRL concepts can be related to the proposed method, they assume the model being imitated exhibits perfect expert behavior. However, in extremely dynamic environments, leveraging a model that exhibits expert behavior is not pragmatic. A policy's behavior is largely unobservable and, therefore, cannot be considered expert behavior for imitation. IRL necessitates traditional model-based reinforcement learning to approximate the reward function, and imitation learning can be similarly approached using traditional methods for deterministic agents, as shown in [43].

Each of the above single-agent flavors of reinforcement learning can be extrapolated to support multi-agent reinforcement learning. Reinforcement learning aligns with olfactory applications because the lack of existing training sets and the dynamism of olfactory navigation will warrant online learning backed by reinforcement learning techniques.

### 10.2 Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) is sometimes referred to as swarm intelligence. While not typically emphasized as a MARL method, [75] demonstrate how a process called co-training can be extended to support MARL, and we elaborate on their methods here. Swarms can be considered as *homogeneous*, where each agent is identical, or *heterogeneous*, where there are many types of different agents within the swarm. Regardless of whether the swarm is homogeneous or heterogeneous, different emergent behaviors arise, and we show how swarm behaviors can be "tuned" based on their heterogeneity, with a focus on the selection of different agents to produce desired collective behavior not unlike the manner in which different activation functions can elicit differing behaviors from a neural network. [200] demonstrate the effectiveness of multiple agents in solving multi-agent games, and we extend much of their philosophy here.

The work of [93] provides an analysis on how deep RL optimizes multi-agent cooperation. They demonstrate the effectiveness of neural networks to enable cooperative information exchange between agents and how global decisions are made as a function of local agent behaviors. [93] focuses on homogeneous swarms where all agents are of the same type. Research from [73] builds off many of their same principles but extend them to investigate heterogeneous swarms and the effects of how different agent policies can calibrate global behavior. [75] provides evidence of swarm collaboration proving especially useful in environments of extreme uncertainty and the estimation of uncertainty is of high importance in plume tracking. Their work draws inspiration [66] and [175] where they emphasize how the evolutionary search space can be factored into logical overlapping subspaces to make optimization easier. Furthermore, [75] and [73] illustrate that even simple tabular policies can provide compelling performance within swarms, as tabular policies limit dimensionality and allow the swarm to build up in complexity rather than the agent.

### 10.3 Co-training and Co-regularization

In 1998, Blum and Mitchell demonstrated the tractability of co-training methodology, assuming that instances from different views are conditionally independent when the co-trainer's classifier makes useful predictions on unlabeled data [23]. Brefeld and Scheffer later improved upon earlier work by Nigam and Ghani by combining a naive Bayes classifier with a support vector machine in their co-training algorithm [27, 141].

Subsequent efforts aimed at developing a more robust co-regularization function were presented by Sindhwani [167] and Wang [189], both of whom attempted to encode prediction dependencies among views into one co-regularization term. However, optimizing the resulting objective function proved challenging and diverged from the core principles of co-training. Most attempts to address the technical complexities of co-training focused on two-view cases and lacked clear performance metrics.

Dai et al. advanced the field by employing pseudo labels and abductive learning to enhance the classification of unlabeled data, incorporating neighboring graphs [50]. Ma et al. introduced SPaCo, which established a more generalizable objective function and self-paced learning technique over a pseudo-supervised co-training algorithm, although it was limited to two-view scenarios [123]. Building on this, Ma et al. integrated additional methods into SPaCo, resulting in SPaMCo, which provided resilience against false negative samples, supported multi-view cases, and improved co-regularization [122]. SPaMCo expanded the applicability of co-training beyond the two-view context.

Following this, Huang et al. demonstrated multi-view co-training in clustering, building on Chang's work [91, 198], while Zheng et al. achieved similar results with image segmentation [203]. Wang et al. introduced the concept of self-paced and self-consistent co-training for image segmentation soon after [187]. Research applying co-training with multi-view or self-paced aspects to swarm intelligence has been limited. Wang et al. explored how two "dueling" models can collaboratively achieve strong performance, but did not utilize co-training, instead suggesting the separation of state-value and action functions [190]. Akella and Lin demonstrated the use of co-training to train a reinforcement learning agent to select actions by learning a temporal policy [5]. Wu et al. proposed using a Q-learning agent to learn a policy for data selection, which then automatically trains co-training classifiers [197]. Here, RL was used to train a supervisor over a classification problem, but the RL-based agent itself was not trained using co-training.

Song et al. showed how co-training can enable an RL agent to learn policies in environments with multiple state-action representations [171]. This work aligns closely with the research by France et al. [75], which combines RL with co-training, but does not include the self-paced element present in our approach. In their work, co-training begins by selecting trajectories that are confidently rewarded according to the optimal policy. Rewarded instances with loss values less than a certain threshold from the observations of each agent are considered confidently rewarded and are selected for the next co-training iteration's training pool. Similarly, unrewarded trajectories are added to the training pool for each agent by sampling values with losses greater than the threshold. When the age parameters are properly tuned and the self-paced training is well-controlled, trajectories selected for the training pool of agent 1 will have a higher probability of being selected for the training pool of agent 2, and vice-versa. This use of co-training with RL agents under extreme ambiguity can be easily mapped to olfaction tasks such as scent-based navigation. Cooperation among multiple robots has proven effective in scent-based navigation [61] and co-training a shared RL policy could add robustness and speed to the task. Since there is little labeled olfaction data, the use of co-training to learn from pseudo-labels and pseudo-rewards seems especially tractable in adaptive learning for scent-based navigation.

### 10.4 Evolutionary Algorithms

Proposed by Marco Dorigo in his 1992 PhD thesis, ACO is another evolutionary optimization technique drawing inspiration from the manner in which ants collaborate to locate a food source. Ants lay pheremones throughout their search in order to direct other ants of the colony to resources within the environment. Each ant represents an agent exploring the search landscape. As each ant of the colony finds a better solution, it lays a "pheremone" signal to direct the search of the others toward the food source, or optimization point. This signal decays as evolution continues to filter out weaker solutions, just as a real pheremone signal decays. In theory, this process continues until a global optimal solution is found.

Most swarms intend to elicit cooperation among all agents within the network to maximize a reward. This intuitively suggests that all agents should be of the same type to facilitate a larger hivemind with a unanimous policy. However, strategically selecting adversarial agents within the swarm can act as a "checks-and-balances" mechanism in order to prevent overfitting of the underlying RL policy. For example, the research of [75] cites that some classes of temporal difference learners are more cooperative in reward-seeking tasks than others. This underscores much of the evidence provided by [133] in how competitive behavior can look similar to cooperative behavior in early stages of games, and how stochastic behaviors can evolve into adversarial ones as games progress. The use of PSO helps generalize all behaviors of the swarm collectively such that all agents are contributing to one single policy.

PSO optimizes a swarm of particles contingent on very simple rules. For each $i^{th}$ particle within the swarm, PSO is optimized according to the following objective function:

$$v_i^{t+1} = wv_i^t + c_1u_1(p_{best}^t - p_i^t) + c_2u_2(p_{global}^t - p_i^t) \quad (7)$$

where $v_i^{t+1}$ is the updated velocity for the $i^{th}$ particle, $w$ is the inertia weight, $v_i^t$ is the particle's current velocity, $c_1$ and $c_2$ are acceleration coefficients, and $u_1$ and $u_2$ respectively denote the cognitive and social weights at time $t$. $g_b^t$ and $p_i^t$ denote the positions of the global best and current $i^{th}$ particle at time $t$ respectively, which in the case of [73] are

**Algorithm 1** Particle Swarm Optimization

Create and initialize each RL agent
**repeat**
  **for** all $\mathbf{p}_i \in \mathbf{P}$ **do**
    Evaluate agent policy fitness $f(\mathbf{p}_i)$
    **if** $f(\mathbf{p}_i) < f(\mathbf{p}_{best})$ **then**
      $\mathbf{p}_{best} = \mathbf{p}_i$
    **end if**
    Evaluate global policy fitness $f(\mathbf{p_{best}})$
    **if** $f(\mathbf{p}_i) < f(\mathbf{p_{global}})$ **then**
      $\mathbf{p_{global}} = \mathbf{p_i}$
    **end if**
    Update velocity for each particle $\mathbf{p}_i$
    $r \leftarrow Random(0, 1)$
    **if** $r <= \sigma_{wind}$ **then**
      $\mathbf{v}_i = 0$
    **end if**
    **if** $r > \sigma_{wind}$ **then**
      $\mathbf{v}_i = w\mathbf{v}_i + c_1u_1(\mathbf{p_{best}} - \mathbf{p_i}) + c_2u_2(\mathbf{p_{global}} - \mathbf{p}_i)$
    **end if**
    $\mathbf{p}_i = \mathbf{p}_i + \mathbf{v}_i$
  **end for**
**until** PSO criterion met or convergence

each the global best RL agent and the current agent being updated.

Algorithm 1 demonstrates the logic behind which the above velocity updates occur. The inertia weight determines the magnitude of change that a particle's previous velocity should have on its updated velocity. The cognitive weight indicates helps regularize each agent's position relative to its previous position and the social weight helps regularize each agent's position globally among the others. $c1$ and $c2$ are commonly referred to as trust parameters [63], where the former expresses how much confidence a particle has in itself and the latter expresses how much confidence the particle has with respect to its neighbors. A variant of parameter-efficient fine-tuning (PEFT) [151] can be employed to find the optimal values for these parameters.

### 10.5 Factoring the Search Space

Factored evolutionary algorithms (FEA) are a type of cooperative co-evolutionary algorithm that form overlapping subpopulations, known as 'factors,' to optimize subsets of variables for a common objective function. These subpopulations act as subproblems of the main optimization function. FEA was formally defined by Strasser et al., who emphasized the importance of selecting an appropriate factor architecture [175]. This definition builds on the original concepts of overlapping swarm intelligence (OSI) introduced by Pillai in [150] and Haberman in [83]. In their 2011 work, Pillai and Sheppard demonstrated the effectiveness of OSI in training artificial neural networks, where each swarm represented a unique path from an input node to an output node. A global perspective of the neural network was maintained using a common vector of weights, which was compiled from the highest fitness particles in each swarm. Our work closely aligns with the OSI model. Pillai and Sheppard also showed that OSI outperformed several alternative cooperative co-evolutionary PSO-based methods and backpropagation. Fortier et al. expanded this by defining Distributed OSI where swarms could communicate values to enhance competition. Butcher et al. built on this by illustrating how information sharing and conflict resolution could be achieved through an actor model, using Pareto improvements via this communication [36, 37]. Both Meerza [127] and Wang [187] applied PSO to reinforcement learning but did not formalize the factored approach. [75] illustrates how to combine co-training and search-space factoring for multi-agent learning.

Factoring of the search space holds promise for swarm intelligence because a factored policy could constrain each agent to optimize a certain sub-task. In olfaction, this could mean one agent is required for surveying one area of the environment for evidence of a target chemical.

### 10.6 Impact of Agent Design on Swarm Functionality

Swarms are a hivemind of their constituent parts, effectively being a behavioral weighted average of all agents within the swarm. However, the intelligence of a swarm can be much greater than the sum of the intelligence of its constituent particles if the swarm is designed correctly. Effective swarm design comes down to effective agent design as well as effectively designing the rules and protocol for inter-agent interaction. Very simple agents working collaboratively can elicit extremely intelligible results.

The 2022 work of Dong, et al. in [59] illustrates this concept well. In their manuscript, they define how simple agents governed by a well-defined protocol can more easily adapt to complex environments than complex agents specially-tuned for different environments. They posit that a practical agent must be resource-conservative and function with bounded memory and computation. Due to this, it becomes infeasible for an agent to maintain a lengthy history of state-action pairs that result in certain rewards. To reconcile this, [59] suggests the use of *agent states* that is leveraged to produce all of its actions. An agent state is an amalgamation of the situational state $S_t$, the epistemic state $P_t$, and the algorithmic state $Z_t$. The situational state is meant to acquire salient information about the agent's current status in its environment. The epistemic state stores information relating to the agent's knowledge of the environment. The algorithmic state records information unrelated to the environment, like time-related metrics from the agent's internal clock or internal random seeds. This can be surmised in the following equation:

$$X_t = (S_t, P_t, Z_t) \tag{8}$$

However, $X_t$ must be updated incrementally since it represents all of the agent's historical knowledge. It can be updated according to the following rule:

$$X_{t+1} = \hat{f}(X_t, A_t, O_{t+1}, U_{t+1}) \tag{9}$$

where $\hat{f}$ represents the agent state update function, $U_{t+1}$ represents algorithmic randomness, $A_t$ represents the action taken at time $t$, and $O_{t+1}$ designates the observation resulting from the action taken at time $t$ subject to the update function and algorithmic randomness at time $t$. At any time $t$, the agent can be seen as executing some policy $\pi_t(|H_t)$ for which action selections depend on the history of agent states $H_t$ conditional on the situational state $S_t$.

With this, the methodology of [59] focuses on evaluating situations subject to their above framework, where rewards are calculated from situational states instead of a bank of historical trajectories or predicted trajectories. The set of situational states is finite which allows bounding of the reward function computation such that it does not require infinite memory. As a result, they posit that the tracking and algorithmic protocol of using situational, epistemic, and algorithmic states enables an agent that can accurately reason about any environment after being initialized according to the following:

1. an initial situational state $S_0 \in S$ and an update function: $f : S\ x\ A\ x\ O \Rightarrow S$
2. a reward function $r : S\ x\ A\ x\ O \Rightarrow [0, 1]$

Dong, et al. emphasize the inherent benefit of leveraging Q-learning and its benefits in constructing the basic agent design based around this framework.

They argue that modern research in Q-learning has merged with literature around regret analysis, mathematical alignment, and provable efficiency, similar to the points [133] discusses in her work around game theory. These optimistic variants of Q-learning agents allow for more efficiency through carefully chosen step sizes and perturbed action value updates that help sustain their optimistic estimations. This leads them to define *discounted Q-learning* for which they leverage in their experiments. They posit that this variant of Q-learning is "the first to establish an algorithm with average regret bounded by a constant multiple of distortion approaches such asymptotic performance within a tractable time frame." In other words, the use of agent states within discounted Q-learning enables the agent to operate indefinitely rather than a predetermined time horizon. The step sizes of the agent are therefore not dependent on the time horizon, a convention that is typical among most reinforcement learning agents during training.

Agent states also effect the ability of the agent to plan. With the use of their framework, the effective planning horizon increases with time and the agent consequently optimizes its performance over arbitrarily long horizons. They note that the effective planning horizon scales with $t^{1/5}$, and this rate leads to their regret bound. This regret bound is dependent on the variability initiated by a fixed situational state update function. In effect, this culminates in the point that restricting the time horizon based on the quantity of data gathered may provide more efficient planning.

Furthermore, the use of agent states enables interactions of simple agents with general environments. The use of a discount factor helps the agent learn more specifically as it explores the environment, although "the discount factor increases over time to generate effective behavior over increasingly long planning horizons.

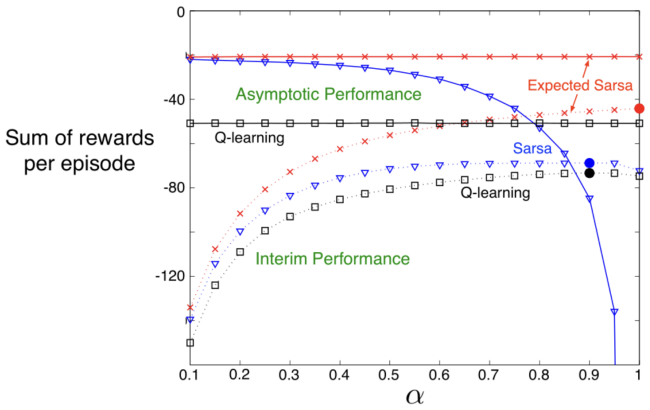


**Figure 14.** Interim and asymptotic performance of temporal difference control on the famous cliff-walking task as a function of learning rate. The solid circles mark the best interim performance of each method. Taken from Sutton & Barto in [177] under the Creative Commons License.

### 10.7 Expected vs Maximum Rewards

Q-learning is an off-policy temporal difference (TD) algorithm that directly approximates the optimal action-value function, regardless of the policy being followed [192]. It learns a policy that selects the next state-action pairs to maximize reward, making it one of the most widely used TD algorithms in reinforcement learning due to its computational efficiency and effective reward approximation. The update rule for each $Q$-value for Q-learning is as follows,

$$Q^*(s,a) = Q(s,a) + \alpha \left[ R(s,a,s') + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \tag{10}$$

Another temporal difference candidate, Expected SARSA, can be tuned as either an on-policy or off-policy method. By design, Expected SARSA works similar to Q-learning, but instead of learning to always greedily select the action that leads to the maximum reward, it learns to always select the action that leads to the *expected* reward. In general, Expected SARSA starts training in a fashion similar to SARSA, where it alters the policy it is learning from as it explores. As training matures and the agent's policy increases in confidence, Expected SARSA can begin building an off-policy

model that, from historic action selections and rewards, always indicates the action that will lead to the average reward among all actions. The update rule for Expected SARSA is shown below.

$$Q^*(s,a) = Q(s,a) + \alpha \left[ R(s,a,s') + \frac{\gamma}{n} \sum_{i=1}^{n} Q(s'_i, a'_i) - Q(s,a) \right] \tag{11}$$

The 2023 study by [75] demonstrates that training a swarm involves co-training a set of policy learners iteratively through the exchange of pseudo-rewarded trajectories. Initially, the rate of pseudo-reward application to trajectories is erratic, resembling random assignment of rewards to state-action pairs. An agent aiming for the average reward can better approximate the hidden true reward of an environment compared to an agent always seeking the maximum reward, like a Q-learner. During the early stages of training, when reward assignment confidence is low, a Q-learner maximizes a lossy policy, whereas an Expected SARSA learner minimizes its losses by approximating the policy. Although regularization terms offer some protection, they are not entirely effective.

In most state-of-the-art reinforcement learning research, where the environment can be directly observed, Q-learning and its variants balance performance and computational cost compared to other TD algorithms. However, based on the evaluation of [75], Q-learner performance depends on the confidence in observing true rewards for actions. If there is high confidence that the observed rewards match the ground truth, a greedy Q-learner is expected to perform better. Therefore, France et al. hypothesize that Expected SARSA may outperform Q-learning in swarm learning, where the Expected SARSA update rule is as follows.

As demonstrated by Ma et al. in their work on image classification [122, 123], and further supported by France [75], confidence in reward assignment improves with successive iterations, leading to more accurate co-training in attributing rewards to state-action trajectories. Consequently, the Expected SARSA policy learner can afford to adopt a more greedy approach. As Sutton and Barto [178] explain, when $\epsilon$ approaches 1, the learner starts selecting actions that yield the highest cumulative reward, effectively transitioning into Q-learning. Expected SARSA leverages this by adjusting policy learning based on correct reward assignment and the self-paced learning parameter $\lambda$. Thus, early iterations of co-training start with a policy resembling SARSA, which mitigates high loss, and gradually transition to Q-learning as the policy becomes more refined. This dual capability allows the algorithm to function as both an on-policy and off-policy learner.

Particularly at the start of co-training, when the pseudo-rewards for unrewarded trajectories are highly inaccurate, Expected SARSA provides smoother learning. Given that the goal is to accurately infer rewards for unobservable states,

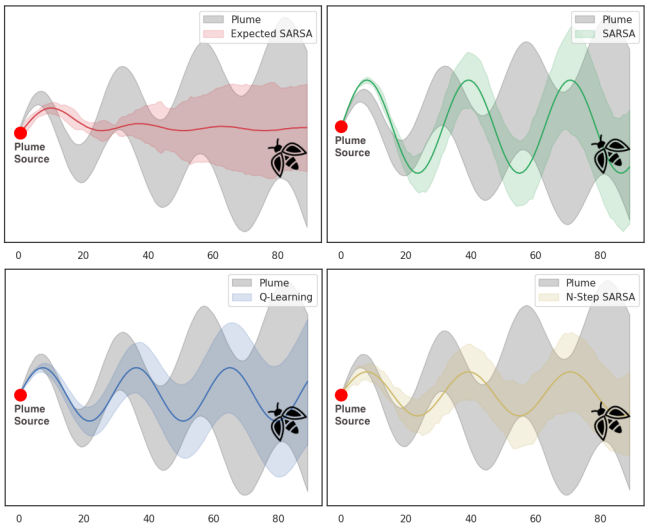


**Figure 15.** From [73], simplified representation of hypothetical plume tracking under different policies. The plume is emitted from left to right, while the agent's progress moves right to left.

the increase in computational complexity might be justified. Expected SARSA allows swarm agents to alternate between on-policy and off-policy methods while naturally controlling reward loss.

### 10.8 Emergent Behaviors in Swarms

One of the blessings of multi-agent learning is also one of the curses: Whilst learning, emerging behaviors arise that can induce alliances, equilibriums, conflicts, and randomness. Intelligent design of the swarm starts with the selection and design of the respective agents. Some behaviors do not manifest until late in learning and training, while others may arise instantly and then diminish asymptotically. As Craig Reynolds famously demonstrated in his famous Boids algorithm [157], simple particles obeying only a handful of very simple rules can result in very intelligent and complex behavior.

Emerging behaviors within swarms is still an under-studied topic. This is partially due to the difficulty in simulating complex multi-agent systems. A direct consequence of this is a demonstration of how agent selection can be used to tune the swarm to favor certain environments. We expect a homogeneous swarm of identical agents to collectively demonstrate behavior analogous to that of their constituent particles. However, this is not always the case, and even with one particle different from the rest of the swarm can elicit unexpected behavior. Figure 15 shows visual representations of how single agents of different policies perform plume tracking. The bold red, green, blue, and yellow lines in each of the figures indicate the optimal path under each agent's policy with confidence bands bounding states that

result from actions each agent would take from the policy it learns. While this is meant only as a visual representation to assess how different each policy is, one can see how each agent's preferences result in different behavioral dynamics. The expected SARSA agent learns to directly navigate to the source through the shortest path overall, but spends significant time outside the plume near the source. The greedy Q-learning agent learns to most accurately follow the turbulence of the plume. One can see that its confidence bands majorly reside within the plume, but the optimal policy line does not indicate the shortest path. The 'n'-step SARSA plot is illustrative of its predictive behavior as the policy center line and confidence bands seem slightly out of phase of the plume turbulence, but becomes less out of phase as the agent moves closer to the source.

The work of [75] emphasizes the importance of evolutionary optimization in their success, and the importance PSO plays in ensuring a globally consistent policy is developed among all agents within the swarm. The underlying behavior of each agent encourages PSO to also resemble performance of other evolutionary techniques similar to, for example, ant colony optimization (ACO) [60]. Throughout their analysis, [73] observe that certain types of agents "lead" the swarm at different stages of evolution. For example, in their experiments, they see SARSA acquiring more rewards during initial evolution iterations due to its affinity for more randomized exploratory action selections in early learning. As evolution progresses, they see Expected SARSA and $n$-step SARSA guiding the swarm due to their ability to filter through noisy states and "look ahead $n$ steps". Finally, as each agent gets closer to the end goal and approaches the final stages of evolution, they observe the greediness of tuned Q-learners being especially helpful in locating the plume source. Impressive analyses from Johanson, et al. in [98] further these notions by demonstrating how agents appear to "barter" to share resources and maximize rewards. Their research emphasizes how both micro- and macro-economic behavior can and should be built into multi-agent systems.

Although these behaviors are not directly programmed into the swarm, they are implicitly learned through what is expected to be due to the effectiveness of PSO optimizing the search landscape as a function of which agents are maximizing the reward signal. In turn, this seems to suggest that, in the context of ACO, each agent type is taking turns guiding the rest of the swarm according to the advantage its reward function provides, which inadvertently resembles pheremone broadcast.

In a similar manner to how agents can be designed to cooperate in a swarm, swarms can be designed to cooperate with other swarms. The same principles can be applied to elicit certain behaviors in swarms of swarms. The factors that influence emergent behavior in single swarms are both compounded and mixed in swarms of swarms, as some of these factors are washed out while others are magnified. For extremely complex or competitive olfactory navigation tasks, multi-swarm cooperation could become an even more important research avenue in olfactory robotics.

### 10.9 Diffusion & Diffusion Policy

A new and very young field of reinforcement learning called *diffusion policy* [41] holds promise for adaptive learning. In [41], they establish a means for which the use of Gaussian noise diffusion can be used to construct a policy for training robots to perform various tasks through reinforcement learning. From an adaptive learning perspective, modeling a reward signal according to noise from the most recent state-action history could be advantageous as the process enables adaptability to new environmental factors. Plume tracking as discussed in Section 6 discusses the intricacies of modeling plumes through different noise conditions. Diffusion policy is attractive here as the diffusion noise could be modeled in proportion to plume dynamics, which may encourage more stable learning. Methods such as DifuzCam from [199] demonstrate how diffusion models can reconstruct high resolution imagery from a minimalist camera. An example for this is demonstrated in Figure 16. Extrapolating these principles to olfaction can enable strong signals to be extracted from high-sensitivity yet high-noise olfaction sensors (like electrochemical sensors). Techniques to reduce the computational cost in diffusion models, such as forward-only diffusion from [121] and flow-matching from [116], could make the idea of diffusion-based methods more attractive for adaptive learning and real-time robotics applications.

The use of policy diffusion in general is young, and its use among olfaction is even more scarce. However, diffusive techniques lend well to various aspects of olfaction such as plume modeling, olfactory navigation, and low-resource learning. The concept of diffusion is not exhaustively covered here, but, as the field of generative AI continues to advance other realms of machine learning, its effect on adaptive learning and machine olfaction is worth mentioning.

## 11 Environment & Simulation

Any sensor-based learning agent needs an environment with which it can interact and freely learn. This process is typically modeled in simulation due to the consequences of leveraging (and crashing) real hardware on prototype models and the freedom that simulations provide to experiment. Simulation is a difficult feat with machine olfaction because, given how young the field is, there are not readily available open-sourced environments that can be simply downloaded and used like there are for other sensor disciplines. Any sensor used in simulation also needs a digital counterpart that functions as close to the real sensor as possible. Many olfactory techniques are new and their digital models must

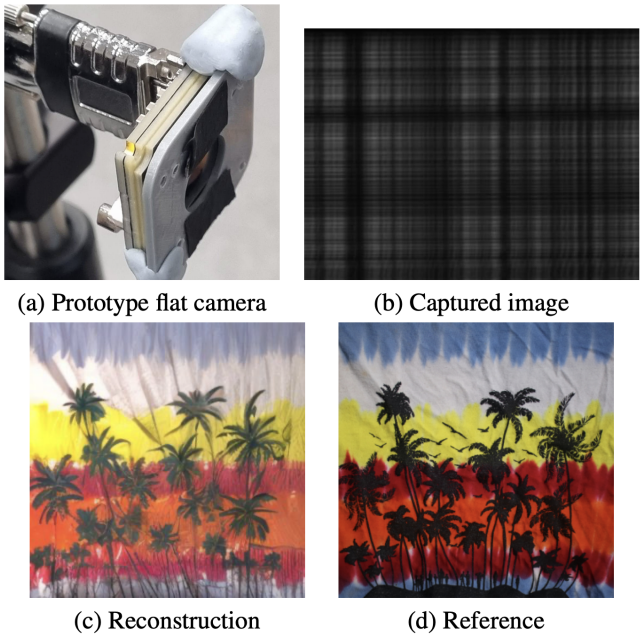


**Figure 16.** Using (a) a prototype flat camera, (b) a measurement image is captured that is not visually understandable. (c) An image is reconstructed from the measurements using text-guided approach. Compare to (d) the reference image captured with a regular camera. See [199] for further details. Taken from Yosef, et al. in [199] under the Creative Commons License.

therefore be constructed before simulation can occur. Environments such as OpenAI's Gym [28] and D4RL [76], and Toyota Research Institute's Drake [41] can be adjusted to accommodate machine olfaction sensors, but a digital replica of the sensor itself must be integrated and the respective frameworks adapted to accommodate. Multi-agent settings are difficult to design due to the inherent computational resources needed and the complexity involved in simulating multiple learners. These difficulties are compounded among extremely dynamic environments, such as olfactory-based plume tracking where air conditions are volatile and the sensor signals are highly sensitive to initial conditions.

### 11.1 "Sim2Real" Gap

No matter how robust the simulation, there is always a "gap" between simulation and reality. Punctually abbreviated as the "sim2real gap", this phenomenon characterizes the differing performance received between simulation and training and the actual performance received when releasing these agents to reality. Nothing has more degrees of freedom than reality, and this problem is exacerbated by the use of machine olfaction where initial conditions, aerodynamics, and part-per-quadrillion chemical changes are more appropriately modeled by chaos theory versus a high degree of freedom simulation engine.

There are several core components to adequately simulating a machine-olfaction based robot. The control loop (or autopilot), learning algorithm, multi-agent interaction, and each individual sensor must all be appropriately modeled. The work of Eschmann, et al. in [64] breaks this problem down well. In their paper, they argue that accurately modeling flight dynamics of a drone can be broken down according to different kinematic derivatives. The complexity and uncertainty in modeling grows as the kinematic derivative grows from position, to velocity, to acceleration, to jerk, snap, crackle, and pop. Eschmann illustrates how the difficulty and uncertainty of modeling should be balanced according to the kinematic access of the simulation and that, to some degree, higher kinematic derivatives may need to be learned due to the uncertainty and non-linearity they provide in simulation and their sensitivity to perturbation in initial conditions. Through the above arguments, they demonstrate the ability for a quadrotor UAV to learn to fly in 18 seconds, and a major contribution to this was a high-fidelity simulator that minimized the delta between simulation and reality. Curriculum learning and a highly optimized simulator enabled shorter reinforcement learning training times that accurately translated to real world control.

In light of the above, of equal importance in simulating any form of sensory control model is the model of the environment. This can prove to be an especially difficult component for machine olfaction, where wind plumes, aerodynamics, and air chemistry highly influence environmental dynamics. The work of [64] is highly influential here because their work shifts focus toward building a more effective simulator while abstracting the controls. They partially argue that the controls and autopilot of the UAV are so non-linear that trying to accurately model them is a fool's errand because it is not possible to construct an accurate model. Too much focus on trying to accurately model such non-linearities can lead to over-optimization of reward signal and therefore erratic learning. These results are compounded with a misrepresentation of the environment. Therefore, they found success by shifting focus in training toward the construction of a higher fidelity environment and abstracting the control model *more*.

With regards to machine olfaction, these findings align with those of [169], where they demonstrate successful olfactory tracking through a rather rudimentary RL model. Both [64] and [169] use actor-critic reinforcement learning with relatively simple neural networks to construct the policy. [169] also emphasizes constructing a robust emulation of environmental conditions that focus on plume dynamics, and note this as a factor for their success in plume tracking.

### 11.2 Environment Design

The work of Samvelyan, et al. in [159] with *MAESTRO* highlights the importance of understanding, not only the agent

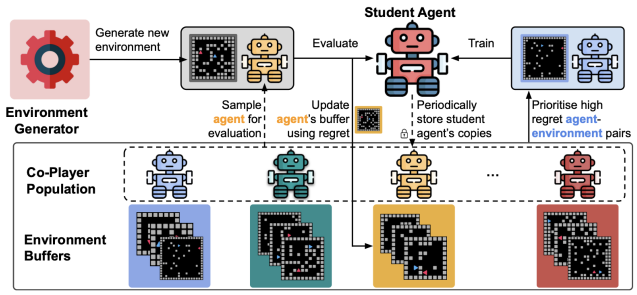


**Figure 17.** MAESTRO keeps a group of co-players, each with its own collection of high-regret environments. As new environments are introduced, the student's regret is measured against the relevant co-player and added to that co-player's collection. MAESTRO consistently supplies high-regret environment/co-player pairs to train the student. Figure adapted from [159] and used under the Creative Commons License.

and its environment disjointly, but their dependencies together. Different equilibria can be established and/or disrupted in multi-agent scenarios depending on if the interactions are inherently cooperative, competitive, stochastic or adversarial. [159] show how environments can be designed more effectively (and autonomously) when these dependencies are known. A diagram illustrating the concept of *MAESTRO* is shown in Figure 17. They argue that the choice of co-players significantly influences the outcome of the game and one way to mitigate this is through self-play. Self-play is a method that allows agents to freely interact toward accomplishing an objective and letting this interaction construct the policy and curriculum for which each agent uses to achieve that objective. Instead of defining the policy beforehand, self-play introduces a *Laissez faire* approach to learning. However, self-play can cause agents to forget how to play against previous versions of their policy due to potential cycles that occur in the strategy space [78]. Self-play should therefore be bound and the authors of [159] describe how to do this through a regret-maximizing teacher that pairs co-players and environments together appropriately.

From a game theory perspective, the objective of the regret-maximizing teacher is to choose an environment and co-player combination for the student. This mention of regret strongly aligns with the work of [133] in multi-agent game theory. When the co-player is disregarded and only the environment with the highest regret is selected, this deduces the selection of the co-player to one that exhibits a suboptimal payoff. Similarly, ignoring the environments and picking the co-player with the highest regret results in an environment that influences a suboptimal policy. This combination is, in turn, suboptimal, with a low level of regret, whereas a teacher considering the joint space can find the optimal co-player environment pair with maximal regret. Consequently, treating the environment and co-player independently can lead to a suboptimal curriculum. This kind of overspecialization towards specific environmental challenges at the cost of overall robustness is often observed in MARL settings [78].

This co-player/environment pairing combined with self-play is one key differentiator of *MAESTRO*. Some works assume that the interactions between agents are known, and therefore do not exhibit self-play. In scenarios where the interactions between agents are not known and must be discovered, or in the event of agent teaming, one must draw more off the work from [133] in being able to identify which agents are cooperative and competitive live. This, again, points back to the importance of continuous learning, which is critical in olfaction scenarios.

The use of *MAESTRO* for environment design for the works of [169] and [49] could expand their results to reflect multi-agent scenarios. Intelligent environment design is crucial for olfaction-based navigation (and any other dynamic sensing scenario).

## 12 Alignment & Ethical Considerations

Historically, AI systems that surpass human performance in narrowly defined domains have frequently led to unforeseen and often detrimental consequences. In computer vision, for instance, facial recognition technologies with superhuman accuracy have raised profound concerns regarding surveillance, privacy violations, and algorithmic bias—especially when deployed without public consent or adequate oversight [33, 155]. Similarly, in natural language processing, large language models have exhibited remarkable fluency, yet simultaneously facilitated the proliferation of misinformation, toxicity, and epistemic distortions at scale [18, 193]. In the domain of audition, AI-generated speech has reached such a degree of realism that it enables the impersonation of individuals with near-perfect fidelity, thereby eroding the distinction between authentic and synthetic media [128].

As we extend artificial intelligence into the realm of olfaction (particularly in pursuit of capabilities that exceed human sensory limits) it is essential to anticipate similar dual-use dilemmas. While superhuman olfactory systems hold promise for beneficial applications such as environmental monitoring, medical diagnostics, and robotic search-and-rescue missions, they may equally be repurposed for invasive monitoring of human physiological states. These include stress levels, fertility status, or substance use—domains of bodily privacy for which current legal and ethical standards offer no clear protection [84]. Just as biometric data like voice and facial features are now recognized as personally identifiable information (PII), emerging research suggests that olfactory signatures, particularly those derived from breath or body odor [14, 15, 100], may soon warrant similar classification.

The ethical complexity deepens when such narrowly superhuman modalities are embedded into unified, multimodal architectures. Recent advances in olfaction-vision-language models (OVLMs) [70, 160] integrate chemical sensing with visual perception and language reasoning, enabling machines to infer behavioral, physiological, and even psychological states of individuals with unprecedented resolution. In adversarial or negligent contexts, biased olfactory agents could mislead users into exposure to harmful substances or fail to detect spoilage or contamination in sensitive domains such as food, pharmaceuticals, or cosmetics. Furthermore, integrating olfactory intelligence into large-scale multimodal foundation models provides machines with yet another sensory axis—granting them increasingly human-like capacities that can be used for surveillance, manipulation, or discriminatory profiling.

The embodiment of olfaction in robotic systems further expands the ways in which AI can explore and interpret the world. This added sensory dimension enhances a robot's ability to form rich multimodal representations of its environment. However, such capabilities also raise serious concerns about covert behavioral inference, commercial exploitation, and the targeting or policing of vulnerable communities. The confluence of modalities does not dilute ethical risk—it amplifies it.

Existing AI ethics frameworks remain largely centered on visual, auditory, and linguistic modalities, with fairness metrics predominantly tailored to social identity attributes such as race, gender, and dialect. These frameworks are ill-prepared to contend with sensory modalities like olfaction, which engage with dimensions of human dignity, consent, and neurophysiological privacy that are less well understood. There is, as yet, no consensus on what constitutes explainability in olfactory AI, nor established methodologies for auditing bias in odor classification systems trained on population-specific scent data.

From an adaptive learning perspective, granting machines the ability to learn on their own could create machines with biases that are difficult to control. Research from [38] and [166] are democratizing the knowledge of embodied AI along with the access to tools and code to build such systems. A proliferation of this knowledge is excellent for the scientific community, but it demands an increased awareness as the above risks become more accessible. We must not only think about what happens when algorithms learn on their own, but what happens when algorithms embodied on robots that interact with the world learn on their own. Methods from federated learning as discussed in Section 8.1 delineate means for controlling updates that are self-learned and contextualizing them against a golden database to prevent rogue behavior. However, any networked system is prone to attack and even a single event of an agent of a federated learning system being compromised and untethered from the master model could result in a catastrophic outcome. Adaptive learning protocols must therefore be extremely robust with remedies for all possible risks considered before deployment.

In the previous sections, several different methods that contribute to adaptive learning in dynamic environments have been discussed, namely multi-modal, multi-agent, and low-resource learning strategies. Each additional modality incorporated into a machine learning model allows for approximation of a larger part of the environment, ruling out more uncertainty. The issues with this arise in finding coupled samples where all modalities are present for the same observation. Low-resource learning strategies and meticulous agent selection can allow the system designer to extract more intelligence from as few of these data points as possible. The use of multiple agents can, like multiple modalities, enable faster approximation of different pockets of the environment that can then be pieced together. Continuous and low-resource learning methods can enable each of these agents to learn with fewer data points more quickly and adapt to rapidly changing environments, a protocol that is mandatory for complex tasks like olfaction-based navigation.

Constructing the optimal reward signal can be difficult, but it is one of the biggest contributors to the success of any reinforcement learning agent. Reinforcement learning with human feedback (RLHF) was developed as a way to help align the reward signal with human preferences. The foundational work was performed by Ouyang, et al. in [144], and now RLHF is heavily used for the training of large language models in responding the correct way. Incorporating humans into the training loop helps to quickly optimize the policy and shape the reward signal among extreme ambiguity. In this sense, RLHF can be heavily beneficial in adaptive learning to help ensure controlled, stable gradients in policy iterations. RLHF is shown to prevent gradient explosion and catastrophic forgetting to promote faster optimization of the reward signal [101]. There is little research that currently exists to support effective strategies for implementing RLHF in active multi-agent learning, but as large language models become increasingly integrated with vision, RLHF is becoming effective in training multi-modal models [101]; we suspect the same can be extrapolated to olfaction models.

As alluded to above, one problem with the use of RLHF for olfaction is the fact that humans cannot verify the presence of many chemical compounds with their own sense of smell. This makes the scope of RLHF limited here, but a derivative of RLHF called RL with AI feedback (RLAIF) could benefit such scenarios. The latter method uses a tuned AI model with several sensors to act as the feedback critic. We expect to see several variants of RLAIF to help advance the state-of-the-art in olfaction models and sensor performance in the future. Strong considerations regarding model alignment will need to be addressed in order to ensure that AI models verifying the truth of olfactory data below human perception is properly accounted for. Little work exists in this area outside the

aroma descriptor studies of the LeffingWell and GoodScents datasets which promotes a fruitful area of research.

## 13 Conclusion

In this survey, we have explored the critical role of adaptive learning in advancing machine olfaction, a field that currently lags behind vision, language, and audio intelligence due to the lack of large, standardized training datasets and consensual benchmarks [74]. Unlike the well-established corpuses for other modalities, olfaction lacks scaled equivalents, which poses significant challenges for achieving state-of-the-art performance. We have highlighted the necessity of active and continuous learning over small datasets to overcome these challenges and achieve notable benchmarks in olfactory tasks such as classification and navigation. By disseminating knowledge about machine olfaction and different gas sensing techniques, we aim to inspire more researchers across all disciplines to engage with and advance the field. Our survey underscores the importance of innovative methods for active, continuous machine learning, which are essential for pushing the boundaries of olfactory robotics and achieving parity with other modalities of AI. We hope this inspires more work in the fields of machine olfaction for AI and robotics.

## Acknowledgments

We would like to thank the faculty of the University of Texas at Dallas Computer Science department for their willingness to support research among students in the doctoral program.

## References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. 2011. Improved algorithms for linear stochastic bandits. In *Proceedings of the Advances in Neural Information Processing Systems*.

[2] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship Learning via Inverse Reinforcement Learning. In *Proceedings of the Twenty-First International Conference on Machine Learning* (Banff, Alberta, Canada) *(ICML '04)*. Association for Computing Machinery, New York, NY, USA, 1.

[3] J. Abernethy, K. A. Lai, and A. Wibisono. 2019. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027* (2019).

[4] A. Agarwal, H. Luo, B. Neyshabur, and R. Schapire. 2017. Corralling a Band of Bandit Algorithms. In *Proceedings of the Conference on Learning Theory*.

[5] Ashlesha Akella and Chin-Teng Lin. 2021. Time and Action Co-Training in Reinforcement Learning Agents. *Frontiers in Control Engineering* 2 (2021).

[6] Ebtsam K. Alenezy, Ahmad E. Kandjani, Mahdokht Shaibani, Adrian Trinchi, Suresh K. Bhargava, Samuel J. Ippolito, and Ylias Sabri. 2025. Human breath analysis; Clinical application and measurement: An overview. *Biosensors and Bioelectronics* 278 (2025), 117094. https://doi.org/10.1016/j.bios.2024.117094

[7] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems* 33 (2020).

[8] Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Ves Stoyanov. 2022. Efficient Large Scale Language Modeling with Mixtures of Experts. arXiv:2112.10684 [cs.CL] https://arxiv.org/abs/2112.10684

[9] Aryballe. n.d.. Automotive Cabin Odors. https://aryballe.com/automotive-cabin-odors/ Accessed: 2025-05-23.

[10] Aryballe. n.d.. DIGITAL OLFACTION & FERMENTATION: WHAT YOU NEED TO KNOW. https://aryballe.com/digital-olfaction-fermentation-what-you-need-to-know/ Accessed: 2025-05-23.

[11] Leffingwell & Associates. [n. d.]. PMP 2001 - Database of perfumery materials and performance. http://www.leffingwell.com/bacispmp.htm. Accessed: 2025-03-08.

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47, 2-3 (2002), 235–256.

[13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal Machine Learning: A Survey and Taxonomy. arXiv:1705.09406 [cs.LG]

[14] Ivneet Banga, Kordel France, Anirban Paul, and Shalini Prasad. 2024. E.Co.Tech Breathalyzer: A Pilot Study of a Non-invasive COVID-19 Diagnostic Tool for Light and Non-smokers. *ACS Measurement Science Au* 4, 5 (16 Oct 2024), 496–503. https://doi.org/10.1021/acsmeasuresciau.4c00020

[15] Ivneet Banga, Anirban Paul, Kordel France, Ben Micklich, Bret Cardwell, Craig Micklich, and Shalini Prasad. 2022. E.Co.Tech-electrochemical handheld breathalyzer COVID sensing technology. *Scientific Reports* 12, 1 (2022), 4370. https://doi.org/10.1038/s41598-022-08321-x

[16] Ahmed Barhoum, Selma Hamimed, Hamda Slimi, Amina Othmani, Fatehy M. Abdel-Haleem, and Mikhael Bechelany. 2023. Modern designs of electrochemical sensor platforms for environmental analyses: Principles, nanofabrication opportunities, and challenges. *Trends in Environmental Analytical Chemistry* 38 (2023), e00199. https://doi.org/10.1016/j.teac.2023.e00199

[17] Leandro Di Bella, Yangxintong Lyu, Bruno Cornelis, and Adrian Munteanu. 2025. HybridTrack: A Hybrid Approach for Robust Multi-Object Tracking. arXiv:2501.01275 [cs.CV] https://arxiv.org/abs/2501.01275

[18] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 610–623.

[19] Christian B. Billesbølle, Claire A. de March, Wijnand J. C. van der Velden, Ning Ma, Jeevan Tewari, Claudia Llinas del Torrent, Linus Li, Bryan Faust, Nagarajan Vaidehi, Hiroaki Matsunami, and Aashish Manglik. 2023. Structural basis of odorant recognition by a human odorant receptor. *Nature* 615, 7953 (01 Mar 2023), 742–749. https://doi.org/10.1038/s41586-023-05798-y

[20] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. 2024. $\pi_0$: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164* (2024).

[21] D. Blackwell. 1956. An Analog of the Minimax Theorem for Vector Payoffs. *Pacific J. Math.* 6, 1 (1956), 1–8.

[22] Eric Block, Seogjoo Jang, Hiroaki Matsunami, Sivakumar Sekharan, Bérénice Dethier, Mehmed Z Ertem, Sivaji Gundala, Yi Pan, Shengju Li, Zhen Li, Stephene N Lodge, Mehmet Ozbil, Huihong Jiang, Sonia F Penalba, Victor S Batista, and Hanyi Zhuang. 2015. Implausibility of the vibrational theory of olfaction. *Proceedings of the National Academy of Sciences of the United States of America* 112, 21 (May 2015),

E2766–E2774. https://doi.org/10.1073/pnas.1503054112 Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S..

[23] Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (Madison, Wisconsin, USA) *(COLT' 98)*. Association for Computing Machinery, New York, NY, USA, 92–100.

[24] Sergei Bobrovnikov, Evgeny Gorlov, and Viktor Zharkov. 2024. Remote Detection and Visualization of Surface Traces of Nitro-Group-Containing Explosives. *Photonics* 11, 11 (2024). https://doi.org/10.3390/photonics11111065

[25] Boston Dynamics. n.d.. Atlas. https://bostondynamics.com/atlas/ Accessed: 2025-05-23.

[26] A. E. Bourgeois and Joanne O. Bourgeois. 2017. Theories of olfaction: A review. *Revista Interamericana de Psicología/Interamerican Journal of Psychology* 4, 1 (Jul. 2017). https://doi.org/10.30849/rip/ijp.v4i1.575

[27] Ulf Brefeld and Tobias Scheffer. 2004. Co-EM Support Vector Learning. In *Proceedings of the Twenty-First International Conference on Machine Learning* (Banff, Alberta, Canada) *(ICML '04)*. Association for Computing Machinery, New York, NY, USA.

[28] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. *CoRR* abs/1606.01540 (2016). arXiv:1606.01540 http://arxiv.org/abs/1606.01540

[29] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *arXiv preprint arXiv:2307.15818*.

[30] Jennifer C Brookes, Filio Hartoutsiou, AP Horsfield, and AM Stoneham. 2007. Could humans recognize odor by phonon assisted tunneling? *Physical review letters* 98, 3 (2007), 038101.

[31] G. W. Brown. 1951. *Iterative solution of games by fictitious play*. Vol. 13. 374–376.

[32] S. Bubeck and A. Slivkins. 2012. The Best of Both Worlds: Stochastic and Adversarial Bandits. In *Proceedings of the Conference on Learning Theory*.

[33] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT)* (2018), 77–91.

[34] Javier Burgués, Victor Hernández, Achim J. Lilienthal, and Santiago Marco. 2019. Smelling Nano Aerial Vehicle for Gas Source Localization and Mapping. *Sensors* 19, 3 (2019). https://doi.org/10.3390/s19030478

[35] Javier Burgués and Santiago Marco. 2020. Environmental chemical sensing using small drones: A review. *Science of The Total Environment* 748 (2020), 141172. https://doi.org/10.1016/j.scitotenv.2020.141172

[36] Stephyn Butcher and John Sheppard. 2018. An Actor Model Implementation of Distributed Factored Evolutionary Algorithms. In *Proceedings of the GECCO Workshop on Parallel and Distributed Evolutionary Inspired Methods* (Kyoto, Japan). 1276–1283.

[37] Stephyn G. W. Butcher, John W. Sheppard, and Shane Strasser. 2018. Information Sharing and Conflict Resolution in Distributed Factored Evolutionary Algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference* (Kyoto, Japan) *(GECCO '18)*. Association for Computing Machinery, New York, NY, USA, 5–12.

[38] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascale, Jade Choghari, Jess Moss, and Thomas Wolf. 2024. LeRobot: State-of-the-art Machine Learning for Real-World Robotics in Pytorch. https://github.com/huggingface/lerobot.

[39] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. 2007. Improved Second-Order Bounds for Prediction with Expert Advice. *Machine Learning* 66, 2-3 (2007), 321–352.

[40] Minqi Chen, Zhongqian Song, Shengjie Liu, Zhenbang Liu, Weiyan Li, Huijun Kong, Cong Li, Yu Bao, Wei Zhang, and Li Niu. 2025. Iontronic tactile sensory system for plant species and growth-stage classification. *Device* 3, 3 (21 Mar 2025). https://doi.org/10.1016/j.device.2024.100615

[41] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. arXiv:2303.04137

[42] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132 [cs.AI]

[43] Kamil Ciosek. 2022. Imitation Learning by Reinforcement Learning. In *International Conference on Learning Representations*.

[44] Clone Robotics. n.d.. Clone Robotics. https://clonerobotics.com Accessed: 2025-05-23.

[45] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, and al. et. 2023. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. https://arxiv.org/abs/2310.08864.

[46] The Good Scents Company. [n. d.]. The Good Scents Company information system. http://www.thegoodscentscompany.com/. Accessed: 2025-03-08.

[47] A. Copper, R. Evand, and M. Ław. 2020. Inductive general game playing. *Machine Learning* 109 (2020), 1393–1434. https://doi.org/10.1007/s10994-019-05843-w

[48] James A. Covington, Santiago Marco, Krishna C. Persaud, Susan S. Schiffman, and H. Troy Nagle. 2021. Artificial Olfaction in the 21st Century. *IEEE Sensors Journal* 21, 11 (1 June 2021), 12969–12990. https://doi.org/10.1109/JSEN.2021.3076412

[49] John Crimaldi, Hong Lei, Andreas Schaefer, Michael Schmuker, Brian H. Smith, Aaron C. True, Justus V. Verhagen, and Jonathan D. Victor. 2022. Active sensing in a dynamic olfactory world. *Journal of Computational Neuroscience* 50, 1 (01 Feb 2022), 1–6. https://doi.org/10.1007/s10827-021-00798-1

[50] Wang-Zhou Dai, Qiuling Xu, Yang Yu, and Zhi-Hua Zhou. 2019. Bridging Machine Learning and Logical Reasoning by Abductive Learning. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc.

[51] Constantinos Daskalakis and Ioannis Panageas. 2020. Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization. *arXiv preprint arXiv:1807.04252v4* (2020).

[52] Antônio J de Miranda-Magalhães, Gustavo M Jantorno, Adauto Z Pralon, Márcio B de Castro, and Cristiano Barros de Melo. 2023. Explosive detection dogs: A perspective from the personality profile, selection, training methods, employment, and performance to mitigate a real threat. *Animals (Basel)* 13, 24 (Dec. 2023), 3773.

[53] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[54] Nik Dennler, Damien Drix, Tom PA Warner, Shavika Rastogi, Cecilia Della Casa, Tobias Ackels, Andreas T Schaefer, André van Schaik, and Michael Schmuker. 2024. High-speed odor sensing using miniaturized electronic nose. *Science Advances* 10, 45 (2024), eadp1764. https://doi.org/10.1126/sciadv.adp1764

[55] Nik Dennler, Shavika Rastogi, Jordi Fonollosa, André van Schaik, and Michael Schmuker. 2022. Drift in a popular metal oxide sensor dataset reveals limitations for gas classification benchmarks. *Sensors and Actuators B: Chemical* 361 (2022), 131668. https://doi.org/10.1016/j.snb.2022.131668

[56] Nik Dennler, Aaron True, André van Schaik, and Michael Schmuker. 2025. Neuromorphic principles for machine olfaction. *Neuromorphic Computing and Engineering* 5, 2 (may 2025), 023001. https://doi.org/10.1088/2634-4386/add0dc

[57] Vikram Narayanan Dhamu, Anirban Paul, Sriram Muthukumar, and Shalini Prasad. 2024. Electrochemical framework for dynamic tracking of Soil Organic Matter. *Biosensors and Bioelectronics: X* 17 (2024), 100440. https://doi.org/10.1016/j.biosx.2024.100440

[58] Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. 2024. Loss of plasticity in deep continual learning. *Nature* 632, 8026 (August 2024), 768–774. https://doi.org/10.1038/s41586-024-07711-7

[59] Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. 2022. Simple agent, complex environment: efficient reinforcement learning with agent states. *J. Mach. Learn. Res.* 23, 1, Article 255 (jan 2022), 54 pages.

[60] Marco Dorigo and Mauro Birattari. 2010. *Ant Colony Optimization*. Springer US, Boston, MA, 36–39. https://doi.org/10.1007/978-0-387-30164-8_22

[61] Bardienus P. Duisterhof, Shushuai Li, Javier Burgués, Vijay Janapa Reddi, and Guido C. H. E. de Croon. 2021. Sniffy Bug: A Fully Autonomous Swarm of Gas-Seeking Nano Quadcopters in Cluttered Environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 9099–9106. https://doi.org/10.1109/IROS51168.2021.9636217

[62] GM Dyson. 1928. Some aspects of the vibration theory of odor. *Perfumery and essential oil record* 19, 456-459 (1928).

[63] Andries P. Engelbrecht. 2007. *Computational Intelligence: An Introduction* (2nd ed.). Wiley Publishing.

[64] Jonas Eschmann, Dario Albani, and Giuseppe Loianno. 2023. Learning to Fly in Seconds. arXiv:2311.13081 [cs.RO]

[65] Figure AI. n.d.. Figure AI. https://www.figure.ai Accessed: 2025-05-23.

[66] Nathan Fortier, John W. Sheppard, and Karthik Ganesan Pillai. 2012. DOSI: Training artificial neural networks using overlapping swarm intelligence with local credit assignment. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*. 1420–1425.

[67] D. Foster, Z. Li, T. Lykouris, K. Sridharan, and E. Tardos. 2016. Learning in Games: Robustness of Fast Convergence. In *Proceedings of the Advances in Neural Information Processing Systems*.

[68] D. J. Foster, S. Kale, M. Mohri, and K. Sridharan. 2017. Parameter-Free Online Learning via Model Selection. In *Advances in Neural Information Processing Systems*. 6022–6032.

[69] D. P. Foster and R. V. Vohra. 1997. Calibrated learning and correlated equilibrium. *Games and Economic Behavior* 21, 1-2 (1997), 40.

[70] Kordel K. France. 2025. Scentience. https://apps.apple.com/us/app/scentience/id6741092923. https://apps.apple.com/us/app/scentience/id6741092923 Accessed: 2025-03-02.

[71] Kordel K. France, Ovidiu Daescu, Anirban Paul, and Shalini Prasad. 2025. Olfactory Inertial Odometry: Sensor Calibration and Drift Compensation. arXiv:2506.04539 [cs.RO] https://arxiv.org/abs/2506.04539

[72] Kordel K. France, Ovidiu Daescu, Anirban Paul, and Shalini Prasad. 2025. Olfactory Inertial Odometry: Sensor Calibration and Drift Compensation. arXiv:2506.04539 [cs.RO] https://arxiv.org/abs/2506.04539

[73] Kordel K. France, Anirban Paul, Ivneet Banga, and Shalini. Prasad. 2024. Emergent Behavior in Evolutionary Swarms for Machine Olfaction. In *Proceedings of the Genetic and Evolutionary Computation Conference* (Melbourne, Australia) *(GECCO '24)*. Association for Computing Machinery, New York, NY, USA, 30–38. https://doi.org/10.1145/3583131.3590376

[74] Kordel K. France, Rohith Peddi, Nik Dennler, and Ovidiu Daescu. 2025. Position: Olfaction Standardization is Essential for the Advancement of Embodied Artificial Intelligence. arXiv:2506.00398 [cs.AI] https://arxiv.org/abs/2506.00398

[75] Kordel K. France and John W. Sheppard. 2023. Factored Particle Swarm Optimization for Policy Co-training in Reinforcement Learning. In *Proceedings of the Genetic and Evolutionary Computation Conference* (Lisbon, Portugal) *(GECCO '23)*. Association for Computing Machinery, New York, NY, USA, 30–38. https://doi.org/10.1145/3583131.3590376

[76] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. arXiv:2004.07219 [cs.LG]

[77] Martin Gardner. 1970. MATHEMATICAL GAMES. *Scientific American* 223, 4 (1970), 120–123. http://www.jstor.org/stable/24927642

[78] Marta Garnelo, Wojciech Marian Czarnecki, Siqi Liu, Dhruva Tirumala, Junhyuk Oh, Gauthier Gidel, Hado van Hasselt, and David Balduzzi. 2021. Pick Your Battles: Interaction Graphs as Population-Level Objectives for Strategic Diversity. *CoRR* abs/2110.04041 (2021). arXiv:2110.04041 https://arxiv.org/abs/2110.04041

[79] Yunhao Ge, Yuecheng Li, Di Wu, Ao Xu, Adam M. Jones, Amanda Sofie Rios, Iordanis Fostiropoulos, Shixian Wen, Po-Hsuan Huang, Zachary William Murdock, Gozde Sahin, Shuo Ni, Kiran Lekkala, Sumedh Anand Sontakke, and Laurent Itti. 2023. Lightweight Learner for Shared Knowledge Lifelong Learning. arXiv:2305.15591 [cs.LG]

[80] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. arXiv:2305.05665 [cs.CV]

[81] Mehmet Gönen and Ethem Alpaydin. 2011. Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research* 12, 64 (2011), 2211–2268. http://jmlr.org/papers/v12/gonen11a.html

[82] Edward W. Graef, Rujuta D. Munje, and Shalini Prasad. 2017. A Robust Electrochemical CO2 Sensor Utilizing Room Temperature Ionic Liquids. *IEEE Transactions on Nanotechnology* 16, 5 (2017), 826–831. https://doi.org/10.1109/TNANO.2017.2672599

[83] Brian K. Haberman and John W. Sheppard. 2012. Overlapping Particle Swarms for Energy-Efficient Routing in Sensor Networks. *Wirel. Netw.* 18, 4 (may 2012), 351–363.

[84] Rosemary Hains, Noah Overman, and Ann-Sophie Barwich. 2021. The ethics of olfactory surveillance: A nascent area of inquiry. *AI & Society* 36 (2021), 841–850.

[85] Elvin Hajizada, Patrick Berggold, Massimiliano Iacono, Arren Glover, and Yulia Sandamirskaya. 2022. Interactive continual learning for robots: a neuromorphic approach. In *Proceedings of the International Conference on Neuromorphic Systems 2022* (Knoxville, TN, USA) *(ICONS '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 10 pages. https://doi.org/10.1145/3546790.3546791

[86] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12. https://doi.org/10.1109/MIS.2009.36

[87] J. Hannan. 1957. Approximation to Bayes Risk in Repeated Play. In *Contributions to the Theory of Games*, Vol. 3. 97–139.

[88] S. Hart. 2005. Adaptive Heuristics. *Econometrica* 73, 5 (2005), 1401–1430.

[89] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, Gabriel Dulac-Arnold, John Agapiou, Joel Z. Leibo, and Audrunas Gruslys. 2018. Deep Q-Learning from Demonstrations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press.

[90] Tajana Horvat, Gordana Pehnec, and Ivana Jakovljević. 2025. Volatile Organic Compounds in Indoor Air: Sampling, Determination, Sources, Health Risk, and Regulatory Insights. *Toxics* 13, 5 (2025). https://doi.org/10.3390/toxics13050344

[91] Zongmo Huang, Yazhou Ren, Xiaorong Pu, Lili Pan, Dezhong Yao, and Guoxian Yu. 2021. Dual self-paced multi-view clustering. *Neural Networks* 140 (2021), 184–192.

[92] Dom Huh and Prasant Mohapatra. 2023. Multi-agent Reinforcement Learning: A Comprehensive Survey. arXiv:2312.10256 [cs.MA]

[93] Maximilian Hüttenrauch, Adrian Šošić, and Gerhard Neumann. 2019. Deep reinforcement learning for swarm systems. *J. Mach. Learn. Res.* 20, 1 (jan 2019), 1966–1996.

[94] M. Hutter and J. Poland. 2005. Adaptive Online Prediction by Following the Perturbed Leader. *Journal of Machine Learning Research* 6, Apr (2005), 639–660.

[95] Nabil Imam and Thomas A. Cleland. 2020. Rapid online learning and robust recall in a neuromorphic olfactory circuit. *Nature Machine Intelligence* 2, 3 (March 2020), 181–191. https://doi.org/10.1038/s42256-020-0159-4

[96] Peter. J Irga, Gabrielle Mullen, Robert Fleck, Stephen Matheson, Sara. J Wilkinson, and Fraser. R Torpy. 2024. Volatile organic compounds emitted by humans indoors– A review on the measurement, test conditions, and analysis techniques. *Building and Environment* 255 (2024), 111442. https://doi.org/10.1016/j.buildenv.2024.111442

[97] Khurram Javed, Haseeb Shah, Richard S. Sutton, and Martha White. 2023. Scalable Real-Time Recurrent Learning Using Columnar-Constructive Networks. *Journal of Machine Learning Research* 24, 256 (2023), 1–34. http://jmlr.org/papers/v24/23-0367.html

[98] Michael Bradley Johanson, Edward Hughes, Finbarr Timbers, and Joel Z. Leibo. 2022. Emergent Bartering Behaviour in Multi-Agent Reinforcement Learning. arXiv:2205.06760 [cs.AI] https://arxiv.org/abs/2205.06760

[99] R. Johari, V. Kamble, and Y. Kanoria. 2016. Matching While Learning. *arXiv preprint arXiv:1603.04549* (2016).

[100] Maria Kaloumenou, Evangelos Skotadis, Nefeli Lagopati, Efstathios Efstathopoulos, and Dimitris Tsoukalas. 2022. Breath Analysis: A Promising Tool for Disease Diagnosis—The Role of Sensors. *Sensors* 22, 3 (2022). https://doi.org/10.3390/s22031238

[101] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A Survey of Reinforcement Learning from Human Feedback. arXiv:2312.14925

[102] Andreas Keller and Leslie B. Vosshall. 2016. Olfactory perception of chemically diverse molecules. *BMC Neuroscience* 17, 1 (08 Aug 2016), 55. https://doi.org/10.1186/s12868-016-0287-2

[103] J. Kennedy and R. Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, Vol. 4. 1942–1948 vol.4. https://doi.org/10.1109/ICNN.1995.488968

[104] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246* (2024).

[105] W. M. Koolen and T. Van Erven. 2015. Second-Order Quantile Methods for Experts and Combinatorial Games. In *Conference on Learning Theory*. 1155–1175.

[106] D. M. Kreps and R. Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27, 2 (1982), 253–279.

[107] Dhireesha Kudithipudi, Catherine Schuman, Craig M. Vineyard, Tej Pandit, Cory Merkel, Rajkumar Kubendran, James B. Aimone, Garrick Orchard, Christian Mayr, Ryad Benosman, Joe Hays, Cliff Young, Chiara Bartolozzi, Amitava Majumdar, Suma George Cardwell, Melika Payvand, Sonia Buckley, Shruti Kulkarni, Hector A. Gonzalez, Gert Cauwenberghs, Chetan Singh Thakur, Anand Subramoney, and Steve Furber. 2025. Neuromorphic computing at scale. *Nature* 637, 8047 (January 2025), 801–812. https://doi.org/10.1038/s41586-024-08253-8

[108] T. Lattimore. 2015. The Pareto Regret Frontier for Bandits. In *Proceedings of the Advances in Neural Information Processing Systems*.

[109] Yann LeCun. 1988. A Theoretical Framework for Back-Propagation. In *Proceedings of the 1988 Connectionist Models Summer School*, David Touretzky, Geoffrey Hinton, and Terrence Sejnowski (Eds.). Morgan Kaufmann, San Mateo, CA, 21–28. https://www.researchgate.net/publication/2360531_A_Theoretical_Framework_for_Back-Propagation

[110] Brian K. Lee, Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie A. Little, Matthew Andres, Britney B. Nguyen, Theresa Moloy, Jacob Yasonik, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, and Alexander B. Wiltschko. 2023. A principal odor map unifies diverse tasks in olfactory perception. *Science* 381, 6661 (2023), 999–1006. https://doi.org/10.1126/science.ade4401 arXiv:https://www.science.org/doi/pdf/10.1126/science.ade4401

[111] K. Leyton-Brown, P. Milgrom, and I. Segal. 2017. Economics and Computer Science of a Radio Spectrum Reallocation. *Proceedings of the National Academy of Sciences* 114, 28 (2017), 7202–7209.

[112] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2024. A Survey on Deep Active Learning: Recent Advances and New Frontiers. arXiv:2405.00334 [cs.LG] https://arxiv.org/abs/2405.00334

[113] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. 2024. A Survey on Benchmarks of Multimodal Large Language Models. arXiv:2408.08632 [cs.CL] https://arxiv.org/abs/2408.08632

[114] Lin Li, Yuze Li, Wei Wei, Yujia Zhang, and Jiye Liang. 2023. Multi-actor mechanism for actor-critic reinforcement learning. *Information Sciences* 647 (2023), 119494. https://doi.org/10.1016/j.ins.2023.119494

[115] Qi Li, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. 2021. Last Iterate Convergence in No-regret Learning: Constrained Minmax Optimization for Convex-concave Landscapes. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, Vol. 130.

[116] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. arXiv:2210.02747 [cs.LG] https://arxiv.org/abs/2210.02747

[117] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

[118] L. T. Liu, H. Mania, and M. Jordan. 2020. Competing Bandits in Matching Markets. In *International Conference on Artificial Intelligence and Statistics*. 1618–1628.

[119] Friedrich Longin, Heiner Beck, Hermann Gütler, Wendelin Heilig, Michael Kleinert, Matthias Rapp, Norman Philipp, Alexander Erban, Dominik Brilhaus, Tabea Mettler-Altmann, and Benjamin Stich. 2020. Aroma and quality of breads baked from old and modern wheat varieties and their prediction from genomic and flour-based metabolite profiles. *Food Research International* 129 (2020), 108748. https://doi.org/10.1016/j.foodres.2019.108748

[120] Stefan Lukow and James C. Weatherall. 2022. Statistical analysis for explosives detection system test and evaluation. *Scientific Reports* 12, 1 (07 Jan 2022), 250. https://doi.org/10.1038/s41598-021-03755-1

[121] Ziwei Luo, Fredrik K. Gustafsson, Jens Sjölund, and Thomas B. Schön. 2025. Forward-only Diffusion Probabilistic Models. arXiv:2505.16733 [cs.LG] https://arxiv.org/abs/2505.16733

[122] Fan Ma, Deyu Meng, Xuanyi Dong, and Yi Yang. 2020. Self-paced Multi-view Co-training. *Journal of Machine Learning Research* 21, 57 (2020), 1–38.

[123] Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. 2017. Self-Paced Co-training. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2275–2284.

[124] G. J. Mailath and L. Samuelson. 2015. *Reputations in Repeated Games*. Vol. 4. Elsevier, 165–238.

[125] G. Malcolm Dyson. 1938. The scientific basis of odour. *Journal of the Society of Chemical Industry* 57, 28 (1938), 647–651. https://doi.org/10.1002/jctb.5000572802

[126] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html

[127] Syed Irfan Ali Meerza, Moinul Islam, and Md. Mohiuddin Uzzal. 2019. Q-Learning Based Particle Swarm Optimization Algorithm for Optimal Path Planning of Swarm of Mobile Robots. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. 1–5.

[128] Thorsten Meyer and Mads Christensen. 2022. Deepfake voice synthesis: A threat to privacy, security, and trust in voice communication. *IEEE Signal Processing Magazine* 39, 1 (2022), 99–106.

[129] Z. Mhammedi, W. M. Koolen, and T. Van Erven. 2019. Lipschitz Adaptivity with Multiple Learning Rates in Online Learning. In *Conference on Learning Theory*. 2490–2511.

[130] P. Milgrom and P. R. Milgrom. 2004. *Putting Auction Theory to Work*. Cambridge University Press.

[131] P. Milgrom and J. Roberts. 1982. Predation, reputation, and entry deterrence. *Journal of Economic Theory* 27, 2 (1982), 280–312.

[132] S. Minton. 1988. Quantitative Results Concerning the Utility of Explanation-Based Learning. In *Proceedings from the Association for the Advancement of Artificial Intelligence*.

[133] Vidya Muthukumar. 2020. *Learning from an unknown environment*. Ph. D. Dissertation. EECS Department, University of California, Berkeley.

[134] Vidya Muthukumar. 2020. *Learning from an Unknown Environment*. Ph. D. Dissertation. University of California at Berkeley.

[135] Vidya Muthukumar, Soham Phade, and Anant Sahai. 2020. On the Impossibility of Convergence of Mixed Strategies with No Regret Learning. *CoRR* abs/2012.02125 (2020). arXiv:2012.02125 https://arxiv.org/abs/2012.02125

[136] V. Muthukumar, M. Ray, A. Sahai, and P. Bartlett. 2019. Best of Many Worlds: Robust Model Selection for Online Supervised Learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*. 3177–3186.

[137] V. Muthukumar and A. Sahai. 2019. Robust commitments and partial reputation. *arXiv preprint arXiv:1905.11555* (2019).

[138] Siddharth Mysore, George Cheng, Yunqi Zhao, Kate Saenko, and Meng Wu. 2022. Multi-Critic Actor Learning: Teaching RL Policies to Act with Style. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJvY_5OzoI

[139] K. Nakamoto. 2023. *Digital Technologies in Olfaction*. Elsevier. https://shop.elsevier.com/books/digital-technologies-in-olfaction/nakamoto/978-0-443-15721-9

[140] N. Newman, A. Fréchette, and K. Leyton-Brown. 2017. Deep Optimization for Spectrum Repacking. *Commun. ACM* 61, 1 (2017), 97–104.

[141] Kamal Nigam and Rayid Ghani. 2000. Analyzing the Effectiveness and Applicability of Co-Training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management* (McLean, Virginia, USA) *(CIKM '00)*. Association for Computing Machinery, New York, NY, USA, 86–93.

[142] Osmo AI. n.d.. Generation by Osmo. https://www.generationbyosmo.com Accessed: 2025-05-23.

[143] Osmo AI. n.d.. Osmo AI. https://www.osmo.ai Accessed: 2025-05-23.

[144] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155

[145] P. Paruchuri, J. P. Pearce, J. Marecki, M. Tambe, F. Ordonez, and S. Kraus. 2008. Playing Games for Security: An Efficient Exact Algorithm for Solving Bayesian Stackelberg Games. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 895–902.

[146] Anirban Paul, Kordel France, Avi Bhatia, Muhanned Abu-Hijleh, Ovidiu Daescu, Ruby Thapa, Rhoda Annoh Gordon, and Shalini Prasad. 2025. Electrochemical breath profiling for early thoracic malignancy screening. *Sensing and Bio-Sensing Research* 49 (2025), 100815. https://doi.org/10.1016/j.sbsr.2025.100815

[147] Anirban Paul, Sriram Muthukumar, and Shalini Prasad. 2019. Review—Room-Temperature Ionic Liquids for Electrochemical Application with Special Focus on Gas Sensors. *Journal of The Electrochemical Society* 167, 3 (dec 2019), 037511. https://doi.org/10.1149/2.0112003JES

[148] T. C. Pearce, S. S. Schiffman, H. T. Nagle, and J. W. Gardner. 2003. *Handbook of Machine Olfaction*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

[149] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelse|a Finn, and Sergey Levine. 2025. FAST: Efficient Action Tokenization for Vision-Language-Action Models. *arXiv preprint arXiv:2501.09747* (2025).

[150] Karthik Ganesan Pillai and John W. Sheppard. 2011. Overlapping swarm intelligence for training artificial neural networks. In *2011 IEEE Symposium on Swarm Intelligence*. 1–8.

[151] Edoardo M. Ponti, Alessandro Sordoni, Yoshua Bengio, and Siva Reddy. 2022. Combining Modular Skills in Multitask Learning. arXiv:2202.13914 [cs.LG]

[152] Tianfei Qi, Ruonan Jia, Linbo Zhang, Guanyu Li, Xuefeng Wei, Jianting Ye, Ruilian Qi, and Haiyuan Chen. 2024. Real-time and visual monitoring coating damage and metal corrosion based on the multiple chromatic transition of TMB triggered by highly active N-CDs/Fe3O4 peroxidase mimics and Fe3+. *Sensors and Actuators B: Chemical* 415 (2024), 136003. https://doi.org/10.1016/j.snb.2024.136003

[153] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021.

Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[154] Bipin Rajendran, Abu Sebastian, Michael Schmuker, Narayan Srinivasa, and Evangelos Eleftheriou. 2019. Low-Power Neuromorphic Hardware for Signal Processing Applications: A Review of Architectural and System-Level Design Approaches. *IEEE Signal Processing Magazine* 36, 6 (2019), 97–110. https://doi.org/10.1109/MSP.2019.2933719

[155] Inioluwa Deborah Raji, Joy Buolamwini, Solon Barocas, Matt Krishnan, and Julian Lemos. 2020. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), 429–435.

[156] Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. The MIT Press.

[157] Craig W. Reynolds. 1998. *Flocks, herds, and schools: a distributed behavioral model.* Association for Computing Machinery, New York, NY, USA, 273–282. https://doi.org/10.1145/280811.281008

[158] S. De Rooij, T. Van Erven, P. D. Grünwald, and W. M. Koolen. 2014. Follow the Leader if You Can, Hedge if You Must. *Journal of Machine Learning Research* 15, 1 (2014), 1281–1316.

[159] Mikayel Samvelyan, Akbir Khan, Michael Dennis, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Roberta Raileanu, and Tim Rocktäschel. 2023. MAESTRO: Open-Ended Environment Design for Multi-Agent Reinforcement Learning. arXiv:2303.03376 [cs.LG]

[160] Scentience. 2025. Scentience App: World's First Olfaction-Vision Language Model. (2025). https://scentience.ai/news/f/scentience-app-worlds-first-olfaction-vision-language-model Accessed: 2025-03-02.

[161] T. C. Schelling. 1980. *The Strategy of Conflict.* Harvard University Press.

[162] Michael Schmuker, Viktor Bahr, and Ramón Huerta. 2016. Exploiting plume structure to decode gas source distance using metal-oxide gas sensors. *Sensors and Actuators B: Chemical* 235 (2016), 636–646. https://doi.org/10.1016/j.snb.2016.05.098

[163] Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. 2024. How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse. arXiv:2404.05090 [cs.LG] https://arxiv.org/abs/2404.05090

[164] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf

[165] Aditi Seshadri, Heta A Gandhi, Geemi P Wellawatte, and Andrew D White. 2022. Why does that molecule smell? (Dec. 2022).

[166] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. 2025. SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics. arXiv:2506.01844 [cs.LG] https://arxiv.org/abs/2506.01844

[167] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. 2005. A Co-Regularization Approach to Semi-supervised Learning with Multiple Views.

[168] Gurpreet Singh, Daniel Lofaro, and Donald Sofge. 2020. Pursuit-evasion with Decentralized Robotic Swarm in Continuous State Space and Action Space via Deep Reinforcement Learning. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART,*. INSTICC, SciTePress, 226–233. https://doi.org/10.5220/0008971502260233

[169] Satpreet H. Singh, Floris van Breugel, Rajesh P. N. Rao, and Bingni W. Brunton. 2023. Emergent behaviour and neural dynamics in artificial agents tracking odour plumes. *Nature Machine Intelligence* 5, 1 (01 Jan 2023), 58–70. https://doi.org/10.1038/s42256-022-00599-w

[170] Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical Networks for Few-shot Learning. arXiv:1703.05175 [cs.LG]

[171] Jialin Song, Ravi Lanka, Yisong Yue, and Masahiro Ono. 2019. Co-training for Policy Learning. *CoRR* abs/1907.04484 (2019). arXiv:1907.04484 http://arxiv.org/abs/1907.04484

[172] S. Sorin. 1999. Merging, Reputation, and Repeated Games with Incomplete Information. *Games and Economic Behavior* 29, 1-2 (1999), 274–308.

[173] H. Von Stackelberg. 1934. *Marktform und gleichgewicht.* J. Springer.

[174] Stanley S. Stevens. 1951. *Handbook of experimental psychology*. Wiley.

[175] Shane Strasser and John W. Sheppard. 2017. Convergence of Factored Evolutionary Algorithms. In *Proceedings of the 14th ACM/SIGEVO Conference on Foundations of Genetic Algorithms* (Copenhagen, Denmark) *(FOGA '17)*. Association for Computing Machinery, New York, NY, USA, 81–94.

[176] Richard S. Sutton. 1988. Learning to predict by the methods of temporal differences. *Machine Learning* 3, 1 (01 Aug 1988), 9–44. https://doi.org/10.1007/BF00115009

[177] R. S. Sutton and A. G. Barto. 2018. *Reinforcement Learning: An Introduction.* MIT Press.

[178] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press.

[179] Thomas Tille. 2012. Automotive suitability of air quality gas sensors. *Sensors and Actuators B: Chemical* 170 (2012), 40–44. https://doi.org/10.1016/j.snb.2010.11.060 Eurosensors XXIV, 2010.

[180] Luca Turin, Simon Gane, Dimitris Georganakis, Klio Maniati, and Efthimios M. C. Skoulakis. 2015. Plausibility of the vibrational theory of olfaction. *Proceedings of the National Academy of Sciences* 112, 25 (2015), E3154–E3154. https://doi.org/10.1073/pnas.1508035112 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1508035112

[181] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs.LG]

[182] T. van Erven and W. M. Koolen. 2016. Metagrad: Multiple Learning Rates in Online Learning. In *Advances in Neural Information Processing Systems*. 3666–3674.

[183] Hado van Hasselt, Sephora Madjiheurem, Matteo Hessel, David Silver, André Barreto, and Diana Borsa. 2021. Expected Eligibility Traces. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 11 (May 2021), 9997–10005. https://doi.org/10.1609/aaai.v35i11.17200

[184] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. 1974. *Theory of Pattern Recognition.* Nauka.

[185] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[186] Alexander Vergara. 2013. Gas sensor arrays in open sampling settings. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5JP5N.

[187] Feng Wang, Xujie Wang, and Shilei Sun. 2022. A Reinforcement Learning Level-Based Particle Swarm Optimization Algorithm for Large-Scale Optimization. *Inf. Sci.* 602, C (jul 2022), 298–312.

[188] Guangzhi Wang, Yuchen Guo, Yang Yu, Yan Shi, Yuxiang Ying, and Hong Men. 2025. ColorNet: An AI-based framework for pork freshness detection using a colorimetric sensor array. *Food Chemistry* 471 (2025), 142794. https://doi.org/10.1016/j.foodchem.2025.142794

[189] Jiao Wang, Siwei Luo, and Yan Li. 2010. A Multi-view Regularization Method for Semi-supervised Learning. In *Advances in Neural Networks - ISNN 2010*, Liqing Zhang, Bao-Liang Lu, and James Kwok (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 444–449.

[190] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. 2016. Dueling Network Architectures for Deep Reinforcement Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 1995–2003.

[191] Tomasz Wasilewski, Jacek Gębicki, and Wojciech Kamysz. 2021. Bio-inspired approaches for explosives detection. *TrAC Trends in Analytical Chemistry* 142 (2021), 116330. https://doi.org/10.1016/j.trac.2021.116330

[192] Christopher Watkins. 1989. *Learning from Delayed Rewards*. Ph. D. Dissertation. Department of Computer Science, Kings College, Cambridge University.

[193] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Christopher Griffin, John Mellor, Po-Sen Huang, Amelia Glaese, Borja Balle, Atoosa Kasirzadeh, Iason Gabriel, et al. 2022. Taxonomy of risks posed by foundation models. *arXiv preprint arXiv:2208.05300* (2022).

[194] Geemi P. Wellawatte and Philippe Schwaller. 2025. Human interpretable structure-property relationships in chemistry using explainable machine learning and large language models. *Communications Chemistry* 8, 1 (14 Jan 2025), 11. https://doi.org/10.1038/s42004-024-01393-y

[195] Wikipedia contributors. n.d.. Optimus (robot) — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Optimus_(robot) Accessed: 2025-05-23.

[196] K. A. Woyach. 2013. *Building Trust into Light-Handed Regulations for Cognitive Radio*. Ph. D. Dissertation. University of California at Berkeley.

[197] Jiawei Wu, Lei Li, and William Yang Wang. 2018. Reinforced Co-Training. *CoRR* abs/1804.06035 (2018). arXiv:1804.06035 http://arxiv.org/abs/1804.06035

[198] Chang Xu, Dacheng Tao, and Chao Xu. 2015. Multi-View Self-Paced Learning for Clustering. In *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina) *(IJCAI'15)*. AAAI Press, 3974–3980.

[199] Erez Yosef and Raja Giryes. 2024. DifuzCam: Replacing Camera Lens with a Mask and a Diffusion Model. arXiv:2408.07541 [cs.CV] https://arxiv.org/abs/2408.07541

[200] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games. arXiv:2103.01955 [cs.LG]

[201] Xueying Zhan, Qingzhong Wang, Kuan hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B. Chan. 2022. A Comparative Survey of Deep Active Learning. arXiv:2203.13450 [cs.LG]

[202] Gengzhi Zhang, Liang Feng, and Yaqing Hou. 2021. Multi-task Actor-Critic with Knowledge Transfer via a Shared Critic. In *Proceedings of The 13th Asian Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 157)*, Vineeth N. Balasubramanian and Ivor Tsang (Eds.). PMLR, 580–593. https://proceedings.mlr.press/v157/zhang21b.html

[203] Xu Zheng, Chong Fu, Haoyu Xie, Jialei Chen, Xingwei Wang, and Chiu-Wing Sham. 2022. Uncertainty-Aware Deep Co-Training for Semi-Supervised Medical Image Segmentation. *Comput. Biol. Med.* 149, C (Oct 2022).

[204] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. 2023. Self-Supervised Multimodal Learning: A Survey. arXiv:2304.01008 [cs.LG]