Towards Safety in Multi-agent Reinforcement Learning through Security and Privacy by Design

Kyle Tilbury
ktilbury@uwaterloo.ca
University of Waterloo

Bailey Kacsmar
kacsmar@ualberta.ca
Amii, University of Alberta

Jesse Hoey
jesse.hoey@uwaterloo.ca
University of Waterloo

Abstract

In multi-agent reinforcement learning (MARL), the integration of security and privacy by design is critical for safe deployment in real-world applications. This position paper explores the unique security and privacy challenges inherent to MARL, identifying potential attack vectors and their implications on system security and user privacy. We emphasize the necessity of embedding security and privacy considerations starting from the initial stages of designing MARL systems, especially in settings involving humans. We highlight theoretical foundations and potential deployment challenges, advocating for a design paradigm that prioritizes security and privacy by design in MARL systems.

1 Introduction

Emerging advances in multi-agent reinforcement learning (MARL) will enable many real-world applications such as collaborative robotics (Wu et al., 2022), autonomous vehicular systems (Bloom et al., 2017; Cai & Xiong, 2023; Pape et al., 2023), and distributed control systems (Wang et al., 2021). The complex and dynamic nature of multi-agent systems, with agents interacting in shared environments that may also be shared by people, raises significant challenges in security and privacy. While recognition of some privacy and security issues afforded by these systems is growing, such as with identifying and mitigating collusion (Foxabbott et al., 2023), verification mechanisms in decentralized settings (Sun et al., 2023), or detecting adversarial attacks (Franzmeyer et al., 2022), the nature and extent of privacy and security vulnerabilities in MARL remain ill-defined.

In this work, we aim to illustrate what characteristics of MARL systems correspond to unique attack vectors, their possible implications, potential defence strategies, and how future MARL systems may better account for these factors in their design, particularly in settings of coordination and cooperation with humans. We outline a variety of attack vectors specific to MARL systems in line with security and privacy work which uses a series of adversarial inference attacks and poisoning attacks as proxies for understanding the privacy and security limits of models (Song & Mittal, 2021; Liu et al., 2022b; Yeom et al., 2018). However, we note that such inference attacks focus on recovering training data or the model parameters; things that do not necessarily translate in a generalized way to the MARL setting. Further, these attacks on machine learning, other than in the case of federated or distributed machine learning are not considering a multi-party setting which is an intrinsic property of MARL. Thus, we bring this conversation to the community to ensure we successfully highlight the necessary MARL system characteristics and foster a conversation on what *privacy by design* can lend to MARL safety as the viability of higher risk applications emerges.

2 Theory to Practice: Challenges from Security and Privacy

Deploying privacy enhancing technologies is non-trivial, even for techniques with privacy formalization work spanning decades (Kacsmar et al., 2020). Achieving formal guarantees of technical privacy requires that we first define what is being protected, from whom, and under what conditions these protections will hold. Further, we must assume that any attacker that targets our system will eventually learn how our system works and the security of our system cannot rely on obfuscation to ensure the protection of our system. This last assumption paraphrases what is known as Kerkhoff's principle and Shannon's Maxim within the security and privacy community (Shannon, 1949).

2.1 Privacy by Design

Given our interest in ensuring the design and deployment of impactful privacy preserving systems for MARL, we turn to the nuanced concept of privacy-by-design (PBD) (Wong & Mulligan, 2019). Essentially, this concept encompasses a codification that to ensure privacy it is required to include privacy from the beginning; before deploying technologies into our digital society. This requirement emerged in recognition of repeated evidence that it is easier and safer to account for privacy and security risks during design rather than having to redesign your system after the fact (Gürses et al., 2011; Atwater et al., 2015). The modern conception of this is attributed to Ann Cavoukian (Cavoukian et al., 2009) and is incorporated into the EU privacy regulation GDPR (e.g., with the property of data minimization, etc.) (Voigt & Von dem Bussche, 2017). Thus, as MARL is currently moving towards application viability, we must begin to assess how to incorporate privacy into the design of applied multi-agent systems.

PBD has a selection of core tenets, including being user centric. However, in the case of MARL, the nature of a user has high variance. As we discuss later within this work, the people impacted by a MARL system could be agents within the system or even just those existing within the environment the MARL system has been deployed to. Further, other tenets of PBD include embedding privacy into the design, having privacy as the default configuration, and ensuring privacy about processes across the whole software life-cycle. For each of these tenets to be upheld, we must first consider how we design MARL systems and where within these different configurations are there greater risks of adversarial action that could infringe upon privacy and security for people in our digital society.

2.2 Privacy and Security Components

In addition to the tenets of PBD, there are a series of vectors that serve as the starting point for the evaluation of privacy and security for multi-agent systems. These vectors include attacker goals, attacker power, and the security and privacy goals. Thus, before defining what 'harms' can occur, we first identify what we want to ensure cannot happen, as well as what must happen. In other words, what functionality do we need and what cost is too high.

Attackers. Within each of the vectors we must consider, there is the notion of an attacker (or adversary). An attacker is an entity that has some targeted goal that is in opposition to the protections we want to preserve within our system. When defining an attacker, we make assumptions about what sorts of computational or informational abilities they have based on their potential 'views' of our system.

An attacker can be a participant in a system or an observer of the system. This means they could even be contributors to training where they have access to a policy, model, training parameters, etc. They can be passive or active. Passive meaning they do not act against the system nor deviate from expected behaviours and active meaning they may take action to cause effects that further their goals.

In some machine learning settings, we can model this attacker similarly to what is done in adversarial learning, however, adversarial learning does not encompass the concept of an attacker more generally nor does it encompass what we mean by attacker within this work (Goodfellow et al., 2020). In the context of MARL, one pre-existing notion of adversaries, is an adversary as some other agent(s) with opposing goals (i.e., opponents) (Littman, 1994). Consider a game-like setting where two agents on opposing teams playing football (soccer) could be thought to be in an adversarial relationship (only one can win the game). A second pre-existing notion of adversaries in MARL is from adversarial training techniques where agents must learn to overcome challenging adverse scenarios or adversarial perturbations to increase their robustness (Pinto et al., 2017). The adversaries we focus on within this work are attackers, aiming to achieve an attacker goal that is distinct from the goals of any participant in the system. That is, returning to the football example, the football agents' purpose is to play against one another. Whereas if our attacker was one of the soccer players, they may be attempting to determine whether the other agent has a particular medical condition based on how they play the game.

Protective Goals. The basic security and privacy principles one starts with are *confidentiality*, *integrity*, and *availability*. We want to ensure that sensitive information is not revealed to unauthorized parties (confidentiality), that data and processes cannot be manipulated without detection (integrity), and that an adversarial actor cannot prevent legitimate actors from accessing and using the system (availability). The system goals aimed at preserving these properties include detecting infringements upon these properties, deterring attackers from infringing upon these properties, preventing attackers from being able to infringe upon these properties, deflect adversaries' attacks to non-critical targets, and recover from any attacks that occur.

Attacker Modelling. The two most common ways of classifying an adversary's actions within a system are honest-but-curious (HBC) or malicious (Evans et al., 2018). In an honest-but-curious setting, the adversary follows all the rules of the protocol, but will try to learn as much as they possibly can from the information they observe. The adversary will not deviate from the protocol in any way and they will not send false information. In contrast to an HBC adversary, a malicious adversary is able to act in ways that deviate from the defined process, including sending on false information, as they work towards succeeding at their attack.

For an adversary, succeeding at their attack corresponds to any of the following goals. The attacker succeeds if they are able to cause loss or harm in some way, are able to intercept something secret, are able to interrupt functionality, are able to modify settings or information from the system, or if they can create their own settings or information in the system without being detected.

In addition to modelling attacker's behaviours and goals, we also need to account for their power and views. For instance, an attacker could have significant monetary, computational, and legal power or the attacker under consideration could be someone with only consumer level computational resources. They may have access to other large datasets that they can use to strengthen their inferences via statistical analysis, or they may only have access to public information such as news articles and social media. Within a multi-agent system, we can define the adversarial view based on whether the adversary has access to parameters, is able to observe the environment state, or even having access to agents' policy information. In other areas of machine learning, whether the adversary has access to such information is typically modelled as white-box access, black-box access, or something in between; with white-box access having the most insight into the system and black-box access having the least (Nasr et al., 2019).

Technical Guarantees. The formal guarantees within security and privacy fall under three possible types. First, we have computational guarantees that make assumptions about the computational hardness of a problem and that an adversary cannot reasonably be expected to have sufficient computational power to break your system. For instance, many technical protocols rely on the assumed mathematical hardness of problems such as the discrete logarithm problem (Goldwasser, 1997). On the opposite side of computational guarantees, we have information theoretic guarantees which rely on systems where no matter how much computational power an adversary has they are unable to violate the security of the system as long as they do not have the secret information, such as a key or collection of shares (Shamir, 1979). Finally, there are statistical guarantees, such as in the case of differential privacy (Dwork, 2006). For a statistical guarantee, the security is framed such that the probability of an adversary being able to discern the information being protected can only occur with a very small probability. In differential privacy, the statistical guarantee can be understood as the probability that an adversary can distinguish between two states, one where a data point contributed to a result and the other a state where the data point did not contribute, is negligibly small. These different types of guarantees can be used in combination to achieve a complete system, though the precise configuration can only be determined once the functionality needs have been formalized, which is why we now turn to the different characteristics of MARL systems.

3 MARL System Characteristics and Adversarial Views

In this section, we conceptualize and contrast MARL system characteristics and discuss them with respect to privacy and security. The characteristics that we will outline are not necessarily exclusive from one another and real-world MARL systems may involve mixtures of some or all of them.

Coordination, Cooperation, and Competition. One vector of significance to any security and privacy analysis is the participant distribution and behaviours; corresponding to the risk of collusion (Blanchard et al., 2017) or even requirements for cooperation (Shamir, 1979). Within MARL, this vector occurs along a spectrum of coordination that spans cooperation to competition. In MARL, *coordination* encompasses the processes through which multiple RL agents in some system act, interact, and, ultimately, achieve some outcome. Multiple agents learn and adapt their policies according to the system's state and other agents in the system. Coordination in MARL often encompasses that agents learn strategies to cooperate, compete, or coexist within their shared environment. Strategies for coordination vary depending on whether agents have aligned, conflicting, or mixed incentives.

In *competitive* MARL settings agents have opposing goals and they aim to maximize their individual welfare (Busoniu et al., 2008). A competitive setting may be structured so that the advantage gained by one agent results in a disadvantage to another. In *cooperative* MARL settings the focus becomes how agents can learn, through their acting in a shared system, to optimize their behaviour such that some measure of global welfare is maximized (Panait & Luke, 2005). Agents can have fully aligned incentives (fully cooperative), fully opposing incentives (fully competitive), or a mixture of both in *mixed* settings. Mixed settings have no restrictions placed upon the goals and relationships among agents and can contain both collaborative and adversarial behaviours (Zhang et al., 2021).

In settings where distributed machine learning is employed (Konečný et al., 2016), the participating entities have a common goal of training a model, but an individual goal of protecting their data. While not having to directly share data is a step towards privacy protections, participants in such schemes have a great deal of insight into the system and know the distribution of their own data. Thus, in such cases there is a requirement for some combination of the following (i) trust in the other participants, (ii) a trusted, or semi-trusted, third-party to facilitate computations (Papernot et al., 2017), or (iii) some computational protections that limit the ability of participants to discern information about one another (Bonawitz et al., 2017). However, unlike distributed machine learning, it is more common within the MARL world to have different notions of the agent's interactions with one another towards the outcome of the system.

Full Autonomy and Mixed Autonomy. While the previous vector illustrated how attackers may have different view points or knowledge depending on the characteristics of the actors in the MARL system, this vector highlights different vulnerabilities. The population of agents present in the system may not be entirely comprised of RL agents. The characteristic of the distribution of RL agents to non-RL agents is classified as either fully autonomous or a mixed-autonomy setting. In a fully autonomous system, all decision making and actions are done by autonomous systems (i.e., RL agents) without direct human intervention. In a mixed-autonomy setting where the environment comprises some mixture of learning agents, other autonomous agents, and human agents (Wu et al., 2017). In the latter, where humans and agents interact and coexist, we have to consider the complexity of social engineering attacks (Mitnick & Simon, 2003; Mouton et al., 2016). In short, the implications of social engineering are that human behaviour can be easier to manipulate and more unpredictable; resulting in exploitation that negatively impact the multi-agent system that human agents and RL agents are operating within.

Even without malicious attackers, human error, where people unintentionally introduce problems into the system through regular interactions without malice can still negatively impact the system. This challenge has resulted in different proposals as to how to formalize human behaviour to mitigate emergent issues due to human action (Ellison, 2007; Basin et al., 2016). Such a formalization requires detailed accounts of different configurations and, thus, are the motivation for our work including an over-view of the MARL system characteristics that may impact security.

Further, the implications of a mixed-autonomy setting are not limited to the security implications introduced by the human involvement. Rather, we also have the potential for privacy harms impacting the human participants, with their involvement creating the possibility of their personal data or information being exposed. Agents learning from human interactions may unintentionally capture (collect, store, make part of policy/model, etc.) private info from those humans without adequate safeguarding of the data; an issue that can only be prevented during design before deployment introduces such a risk.

Training and Execution. One of the more unqiue characteristics of MARL, at least with respect to security and privacy, is the categories of training and execution. MARL learning broadly falls into three categories: centralized training and execution, centralized training and decentralized execution, and decentralized training and execution (Albrecht et al., 2024). *Centralized training and execution* is where both the training process and decision-making during execution are managed centrally. This means that all agents have some shared type of information or mechanism between them, such as a centralized controller that can access the full state of the environment and access the actions of all agents. *Centralized training and decentralized execution* constitutes the paradigm where agents have some shared information or mechanism during training, but act independently based only on their own observations during testing or deployment. *Decentralized training and execution* encapsulates scenarios where both the learning and deployment are performed independently by each agent without any form of centralized control or shared information between them.

These configurations are distinct from other types of computation with centralization, such as federated learning (Konečnỳ et al., 2016). Centralization in MARL has greater control and knowledge than other notions of centralization. Further, in some applications, it may be crucial to keep certain information private to an agent and centralized mechanisms that share information could unwittingly make it easier for attackers to access privileged information. Work will be needed to analyze what protections are needed to prevent centralization from introducing risks rather than mitigating them. That is, while a centralized setting could be modelled as only having a single entity an outside attacker could target, if the centralization leaks information to agents within the system, there could be many targets. Alternatively, an attacker that successfully compromises the central controller would compromise the entire system. Thus, it is critical that work moves towards elucidating the possible amplifications or unique attack vectors afforded to attackers by these three paradigms. In particular these paradigms do not map to existing security and privacy work on machine learning, so we cannot hypothesize the risk bounds for MARL systems based on existing attacks.

Communication. One of the vectors that heavily influences security and privacy attacks is what information is available to the attacker, both inside and outside of the system. The communication paradigms within MARL are quite diverse and facilitate information sharing among agents (Zhu et al., 2024). There are settings with no communication, where agents act independently as well as settings with structured forms of communication. For example, broadcast type communication allows agents to send messages that are received broadly by all other agents in the environment (Foerster et al., 2016). Directed communication allows targeted messaging between specific agents (Ding et al., 2020). Additionally, implicit communication involves agents inferring information from actions they observe of other agents (Tian et al., 2020). With respect to communication, this MARL characteristic will impact what an attacker can know inside of the system, corresponding to different settings similar to what is known as white-box and black-box attacks (Nasr et al., 2019).

Online and Offline Learning. One of the most apparent vectors that can influence what attacks are feasible, is whether the MARL system has agents learning online, offline, or some combination. In *online learning*, within the context of MARL, agents learn and update their policies while actively gathering experience the environment. Agents continuously update their strategies based on the feedback, i.e. rewards and states, received from the environment in real-time. Whereas in *offline learning*, agents are trained on a fixed dataset of experiences, typically without further interaction with the environment during initial learning (Yang et al., 2021; Meng et al., 2023). Datasets used for offline learning are typically collected from previous experiences of agents gathered in similar environments. In a combination setting, agents may be pre-trained in an offline manner and then undergo a period of online learning whether during further training or during deployment.

The security and privacy distinction along this vector corresponds to the distinction between active and passive adversaries in secure multi-party computation (Evans et al., 2018). In offline learning, an adversary will not be able to influence the system or observe changes to policies that occur based on their actions. This adversary will only observe actions and outputs to make hypotheses. On the opposite end of this spectrum, in an online setting an adversary could take actions to change the environment in clever ways to impact the agents' policies and what actions they take. In short, offline learning is primarily at risk of passive attacks from adversaries while online learning must also face active attackers. However, while active adversaries are possible if they are already able to exist within and influence the environment, our next vector has the risk of active adversaries joining their environment as well.

Continual Learning. Continual lifelong learning is when agents continue to learn and adapt their policies throughout their deployment (Al-Shedivat et al., 2017). Continual learning agents may be more adaptable, which may be attributed to their ability to adjust their strategies to new information or changes in the environment. However, such adaptations could result in unpredictable or unstable behaviour when the changes to the environment and new information come from an attacker. Further, in this setting an attacker could make multiple entities to influence the environment on their behalf to more efficiently corrupt the system (Douceur, 2002). In contrast to the context of continual learning, fixed learning limits agents' learning to a defined training phase after which the agents' policies are no longer updated, not even during deployment or execution. Employing a fixed setting in safety-critical applications may be advisable as the policies are stable and consistent and cannot be influenced by unforeseen attacker actions. In summary, the setting for continual or fixed learning directly corresponds to the viability of an attacker executing what are referred to as poisoning attacks on other forms of machine learning (Wang et al., 2022).

The Greater MARL Pipeline. Even beyond the characteristics we have discussed thus far, MARL is not monolithic. MARL can use different types of RL, but also supervised and unsupervised machine learning. At the core, different reinforcement learning methods impact privacy and security. Consider model-free approaches versus model-based approaches (Sutton & Barto, 2018). The model of the environment built in model-based approaches could encode sensitive personal data which could be more easily exploited by an attacker than a model-free approach where an attacker might have to rely on an inference attack to glean sensitive information about the training data used. Further, other aspects such as RL algorithm selection could be problematic if some algorithms are more prone to overfitting to some data or patterns during training, which is already known to lead to more successful inference attacks for other machine learning systems (Yeom et al., 2018).

Incorporating supervised and unsupervised learning methods into multi-agent systems will correspond to introducing their known attack vulnerabilities, with many attacks having already been developed (Papernot et al., 2016; 2017; Nasr et al., 2019; Carlini et al., 2021) Some examples where supervised learning can be leveraged in MARL include supervised pretraining (Schwarzer et al., 2021; Lee et al., 2024), supervised policy distillation (Wadhwania et al., 2019), or imitation learning (Song et al., 2018). Unsupervised learning in MARL can be utilized for things such as representation learning (Grover et al., 2018; Laskin et al., 2020), generating intrinsic motivation (Bellemare et al., 2016), or self-supervised learning (Pathak et al., 2017). While incorporating these other ML approaches can enhance MARL systems, they facilitate the need to incorporate the broader knowledge of privacy and security in machine learning (Papernot et al., 2016).

4 Discussion

While work has emerged towards improving the safety of reinforcement learning (Melcer et al., 2022; Marchesini et al., 2023; Marzari et al., 2023; Melcer et al., 2024), recent work has only begun to highlight the need for some notion of balance with respect to learning retention versus the need for privacy and security through consideration of unlearning in the RL domain (Liu et al., 2022a). We want to emphasize the following three security and privacy components that need to be addressed by both the MARL research community and the security and privacy research community.

First, formalization of attackers for MARL deployments. Across the characteristics we describe above, any number of combinations exist that can amplify or mitigate risk. Possible adversarial abilities, resources, as well as knowledge are dependent on whether an attacker has access to a central or singular view point, whether an attacker can generate its own agents it controls, whether social engineering of human actors can aid in their attack, and even whether they can influence the environment in the case of continuous or certain online learning settings.

Second, we must consider the possible adversarial goals and their corresponding attacks. An attacker may aim to control an agent, and thus their actions, aim to influence the selection of an agents' policies, impact the environment the agents act within, and so on towards any number of outcomes including city gridlock or disrupting water access depending on the application under consideration. There is currently potential for attackers to impact decision making in MARL without the attackers' actions being detectable (Franzmeyer et al., 2022). This existing attack can be executed within environments where both the attacker and the victim are agents that interact, showing that there is already risk along two of the vectors we highlight as factors in our overview of MARL system characteristics. These attacks are designed to manipulate the observations (and thus decision) of the victim agent without being detectable; and thus, hindering current designs from being able to recover from such an attack.

Finally, we must move towards incorporating defences within the design of MARL systems. Further, defences cannot be limited to data collection concerns, which while important are not the full extent of the problem to be addressed. While recent work investigates what information is strategically relevant and irrelevant in cooperative multi-agent environments (Lauffer et al., 2023) it is only the beginning. We need to incorporate techniques to detect attacker transgressions. For instance, within a continuous learning setting, if we are able to detect malicious actors (by bounding normal behaviour via some form of interpretability) we may be able to detect misbehaving agents or anomalies within the world and subsequently incorporate "mental state" check points, policy check points that allow us to go back to a recovery state from before an adversary either corrupted the world or influenced the agents' actions.

We want to engage the MARL community at this workshop and also the security community in the future. The challenges for deploying any form of machine learning when considering security and privacy are substantial. They include compliance with emerging legal regulations in the EU (Edwards, 2021), Canada (Government of Canada, 2022), and other governing bodies in addition to the technical challenges. The path towards overcoming these challenges and achieving deployment in high-risk settings requires incorporating privacy and security into the design of MARL now.

References

Maruan Al-Shedivat, Trapit Bansal, Yuri Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous Adaptation via Meta-Learning in Nonstationary and Competitive Environments. *arXiv preprint arXiv:1710.03641*, 2017.

Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024.

Erinn Atwater, Cecylia Bocovich, Urs Hengartner, Ed Lank, and Ian Goldberg. Leading Johnny to Water: Designing for Usability and Trust. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pp. 69–88, Ottawa, July 2015. USENIX Association. ISBN 978-1-931971-249.

David Basin, Saa Radomirovic, and Lara Schmid. Modeling Human Errors in Security Protocols. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pp. 325–340. IEEE, 2016.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying Count-Based Exploration and Intrinsic Motivation. *Advances in Neural Information Processing Systems*, 29, 2016.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Cara Bloom, Joshua Tan, Javed Ramjohn, and Lujo Bauer. Self-Driving Cars and Data Collection: Privacy Perceptions of Networked Autonomous Vehicles. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pp. 357–375, Santa Clara, CA, July 2017. USENIX Association. ISBN 978-1-931971-39-3.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191. ACM, 2017.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Zekun Cai and Aiping Xiong. Understand Users' Privacy Perception and Decision of V2X Communication in Connected Autonomous Vehicles. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2975–2992, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Ann Cavoukian et al. Privacy by Design: The 7 Foundational Principles. *Information and privacy commissioner of Ontario, Canada*, 5:12, 2009.

Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning Individually Inferred Communication for Multi-Agent Cooperation. *Advances in Neural Information Processing Systems*, 33:22069–22079, 2020.

John R Douceur. The Sybil Attack. In *International workshop on peer-to-peer systems*, pp. 251–260. Springer, 2002.

Cynthia Dwork. Differential Privacy. In *International Colloquium on Automata, Languages, and Programming*, pp. 1–12. Springer, 2006.

Lilian Edwards. The EU AI Act: A Summary of its Significance and Scope. *Artificial Intelligence (the EU AI Act)*, 1, 2021.

Carl M. Ellison. Ceremony Design and Analysis. *IACR Cryptology ePrint Archive*, 2007:399, 2007.

David Evans, Vladimir Kolesnikov, Mike Rosulek, et al. A Pragmatic Introduction to Secure Multi-Party Computation. *Foundations and Trends® in Privacy and Security*, 2(2-3):70–246, 2018.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Jack Foxabbott, Sam Deverett, Kaspar Senft, Samuel Dower, and Lewis Hammond. Defining and Mitigating Collusion in Multi-Agent Systems. In *Multi-Agent Security Workshop at NeurIPS*, 2023.

Tim Franzmeyer, Stephen McAleer, João F Henriques, Jakob N Foerster, Philip HS Torr, Adel Bibi, and Christian Schroeder de Witt. Illusory Attacks: Detectability Matters in Adversarial Attacks on Sequential Decision-Makers. *arXiv preprint arXiv:2207.10170*, 2022.

Shafi Goldwasser. New Directions in Cryptography: Twenty Some Years Later (or Cryptograpy and Complexity Theory: A Match Made in Heaven). In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pp. 314–324. IEEE, 1997.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144, 2020.

Government of Canada. Bill C-27: An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts. https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html, 2022.

Aditya Grover, Maruan Al-Shedivat, Jayesh Gupta, Yuri Burda, and Harrison Edwards. Learning Policy Representations in Multiagent Systems. In *International Conference on Machine Learning*, pp. 1802–1811. PMLR, 2018.

Seda Gürses, Carmela Troncoso, and Claudia Diaz. Engineering Privacy by Design. *Computers, Privacy & Data Protection*, 14(3):25, 2011.

Bailey Kacsmar, Chelsea H Komlo, Florian Kerschbaum, and Ian Goldberg. Mind the Gap: Ceremonies for Applied Secret Sharing. *Proceedings on Privacy Enhancing Technologies*, 2020.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. In *International Conference on Machine Learning*, pp. 5639–5650. PMLR, 2020.

Niklas Lauffer, Ameesh Shah, Micah Carroll, Michael D Dennis, and Stuart Russell. Who needs to know? Minimal Knowledge for Optimal Coordination. In *International Conference on Machine Learning*, pp. 18599–18613. PMLR, 2023.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised Pretraining Can Learn In-Context Reinforcement Learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Michael L Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Machine Learning Proceedings 1994*, pp. 157–163. Elsevier, 1994.

Bo Liu, Qiang Liu, and Peter Stone. Continual Learning and Private Unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022a.

Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA, August 2022b. USENIX Association.

Enrico Marchesini, Luca Marzari, Alessandro Farinelli, and Christopher Amato. Safe Deep Reinforcement Learning by Verifying Task-Level Properties. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1466–1475, 2023.

Luca Marzari, Enrico Marchesini, and Alessandro Farinelli. Online Safety Property Collection and Refinement for Safe Deep Reinforcement Learning in Mapless Navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7133–7139. IEEE, 2023.

Daniel Melcer, Christopher Amato, and Stavros Tripakis. Shield Decentralization for Safe Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35:13367–13379, 2022.

Daniel Melcer, Christopher Amato, and Stavros Tripakis. Shield Decentralization for Safe Reinforcement Learning in General Partially Observable Multi-Agent Environments. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2384–2386, 2024.

Linghui Meng, Muning Wen, Chenyang Le, Xiyun Li, Dengpeng Xing, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, Yaodong Yang, et al. Offline Pre-Trained Multi-Agent Decision Transformer. *Machine Intelligence Research*, 20(2):233–248, 2023.

Kevin D Mitnick and William L Simon. *The Art of Deception: Controlling the Human Element of Security*. John Wiley & Sons, 2003.

Francois Mouton, Louise Leenen, and Hein S Venter. Social Engineering Attack Examples, Templates and Scenarios. *Computers & Security*, 59:186–209, 2016.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks Against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 739–753. IEEE, 2019.

Liviu Panait and Sean Luke. Cooperative Multi-Agent Learning: The State of the Art. *Autonomous agents and multi-agent systems*, 11:387–434, 2005.

Sebastian Pape, Sarah Syed-Winkler, Armando Miguel Garcia, Badreddine Chah, Anis Bkakria, Matthias Hiller, Tobias Walcher, Alexandre Lombard, Abdeljalil Abbas-Turki, and Reda Yaich. A systematic approach for automotive privacy management. In *Proceedings of the 7th ACM Computer Science in Cars Symposium*, CSCS '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400704543. doi: 10.1145/3631204.3631863. URL https://doi.org/10.1145/3631204.3631863.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the Science of Security and Privacy in Machine Learning. *arXiv preprint arXiv:1611.03814*, 2016.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. In *the International Conference on Learning Representations*, Toulon, France, 2017.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-Driven Exploration by Self-Supervised Prediction. In *International Conference on Machine Learning*, pp. 2778–2787. PMLR, 2017.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust Adversarial Reinforcement Learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R Devon Hjelm, Philip Bachman, and Aaron C Courville. Pretraining Representations for Data-Efficient Reinforcement Learning. *Advances in Neural Information Processing Systems*, 34:12686–12699, 2021.

Adi Shamir. How to Share a Secret. *Communications of the ACM*, 22(11):612–613, 1979.

Claude E Shannon. Communication Theory of Secrecy Systems. *The Bell System Technical Journal*, 28(4):656–715, 1949.

Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-Agent Generative Adversarial Imitation Learning. *Advances in neural information processing systems*, 31, 2018.

Liwei Song and Prateek Mittal. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2615–2632, 2021.

Xinyuan Sun, Davide Crapis, Matt Stephenson, Barnabé Monnot, Thomas Thiery, and Jonathan Passerat-Palmbach. Cooperative AI via Decentralized Commitment Devices. *arXiv preprint arXiv:2311.07815*, 2023.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction.* MIT press, 2018.

Zheng Tian, Shihao Zou, Ian Davies, Tim Warr, Lisheng Wu, Haitham Bou Ammar, and Jun Wang. Learning to Communicate Implicitly by Actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7261–7268, 2020.

Paul Voigt and Axel Von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide.* Springer, 2017.

Samir Wadhwania, Dong-Ki Kim, Shayegan Omidshafiei, and Jonathan P How. Policy Distillation and Value Matching in Multiagent Reinforcement Learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8193–8200. IEEE, 2019.

Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-Agent Reinforcement Learning for Active Voltage Control on Power Distribution Networks. *Advances in Neural Information Processing Systems*, 34:3271–3284, 2021.

Zhibo Wang, Jingjing Ma, Xue Wang, Jiahui Hu, Zhan Qin, and Kui Ren. Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems. *ACM Computing Surveys*, 55(7):1–36, 2022.

Richmond Y Wong and Deirdre K Mulligan. Bringing Design to the Privacy Table: Broadening “design” in “privacy by design” Through the Lens of HCI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–17, 2019.

Cathy Wu, Aboudy Kreidieh, Eugene Vinitsky, and Alexandre M Bayen. Emergent Behaviors in Mixed-Autonomy Traffic. In *Conference on Robot Learning*, pp. 398–407. PMLR, 2017.

Ruofan Wu, Junmin Zhong, Brent Wallace, Xiang Gao, He Huang, and Jennie Si. Human-Robotic Prosthesis as Collaborating Agents for Symmetrical Walking. *Advances in Neural Information Processing Systems*, 35:27306–27320, 2022.

Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe What You See: Implicit Constraint Approach for Offline Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems*, 34:10299–10312, 2021.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):4, 2024.