Milnor-Myerson Games and The Principles of Artificial Principal-Agent Problems

Manfred Diaz
diazcabm@mila.quebec
Mila, Université of Montréal

Joel Z. Leibo
jzl@deepmind.com
Google DeepMind

Liam Paull
paulll@mila.quebec
Mila, Université de Montréal

Abstract

In this paper, we introduce Milnor-Myerson games, a multiplayer interaction structure at the core of machine learning (ML), to shed light on the fundamental principles and implications the artificial principal-agent problem has had in landmark ML results like AlphaGo and large language models (LLMs).

1 Introduction

Since its inception, artificial intelligence (AI) has studied the construction of *artificial* agents that can think and act like humans (Turing, 1950; McCarthy et al., 1955; Russell & Norvig, 2020). For almost two decades, and spearheaded by notable results in reinforcement learning (Silver et al., 2016; 2018; Vinyals et al., 2019; Berner et al., 2019) and, more recently, in (self) supervised learning at scale (OpenAI, 2023; Gemini Team, 2023), machine learning (ML) has positioned itself as the *de facto* approach to produce agents with those capabilities. However, several important questions still linger. Contrast, for instance, the outstanding *go*-playing abilities of *AlphaGo* (Silver et al., 2016) and the natural language understanding capacities of Large Language Models (LLM) (OpenAI, 2023; Gemini Team, 2023). In *AlphaGo*, we found an artificial agent that challenged more than two millennia of human go-playing knowledge with Move 37[1]. In current LLMs, we find general-purpose agents with a remarkable ability to comprehend human-human and human-world interactions through natural language (Bubeck et al., 2023). In this paper, we argue that a thread connecting these two outstanding results is baked into a multiplayer structure that exists at the core of ML problems.

To this end, we note that reinforcement learning (Sutton & Barto, 2018) and supervised learning (Bishop, 2006; Goodfellow et al., 2016) problems share a common construction. In each, there are at least three functionally equivalent components, namely, a *model* or *policy*, a *data distribution* or *transition dynamics*, and an *accuracy* or *reward* function. The meaning of these three components hides a multiplayer structure. The first of these players, the *principal*, arises from a delegation problem (Ross, 1973). If we consider an artificial agent to be an entity that would act on a human's behalf, problem designers have entered into a principal-agent relationship with artificial agents, and played the role of the *principal*, a player that defines the payoff of other players, the *agents*, to guarantee a specific outcome for herself or others (Myerson, 1983). In ML, problem designers construct *reward* or *accuracy* functions, the *principal strategy*, to elicit agent behavior consistent with their preferences. A second player, *nature*, emerges by noting that when a (decision-making) player aims to maximize its payoff in a two-player game against an opponent that plays with a fixed strategy and constant payoff, it plays a game against nature (Milnor, 1951). The *data distribution* in supervised and the *transition dynamics* in reinforcement learning sketch *nature's strategy*, and model a range of world-grounded problems (Russakovsky et al., 2014; Brockman et al., 2016; Radford et al., 2019).

[1] Move 37 refers to the 37th move that *AlphaGo* produced during the second match of the six-match series against *go* world-champion Lee Sedol (Menick, 2016; Kohs, 2017).

Therefore, ML problems share this three-player structure that we introduce in Sec. 3, called *Milnor-Myerson games*[2], where human designers, acting as *principals*, encode in their strategies their preferences for the behaviour an agent uses in its game against the sketch of *nature*'s strategy the designers provide. This structure extends to other areas of ML by considering the interactions among multiple agents and multiple principals. These higher-order structures present an opportunity to explore the human designer-artificial agent relationship. In Sec. 4, we develop formal notions of the artificial principal-agent problem, explain the modes of principal supervision and how they connect to sparse and dense objectives, and describe how ML algorithms function as propagation mechanisms for the principal's preferences. Later, in Sec. 5, we leverage these formalisms to connect *AlphaGo* and LLMs. The *Principal's Principle of Indifference*, a proposition that we derived from the observed bounded rationality found in human principals (Simon, 1990; Selten, 1990) and the biases strong forms of supervision have introduced in ML systems (Silver et al., 2017; Wang et al., 2024), advocates that, in the road towards artificial agents that can think and act as humans and beyond, whenever human designers are uncertain about the consequences of their preferences, they should refrain for expressing them.

2 Background

Games Against Nature. A game against nature is an asymmetric two-player game that we denote by $\mathcal{G}_\rho = \langle \mathcal{X}, \mathcal{Y}, \ell \rangle$ that a player, the *decision-making player*, with action space $\mathcal{Y}$, strategy $\pi \in \Delta(\mathcal{Y})$, and payoff function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathrm{R}$, plays against *nature*, a player with action space $\mathcal{X}$, strategy $\rho \in \Delta(\mathcal{X})$, and whose payoff is unknown or identically zero (Milnor, 1951). If nature's strategy is assumed known, the agent solves a problem of *decision making under risk* (Duncan Luce & Raiffa, 1989; Peterson, 2017), and solutions can be derived from game theory and the *expected utility* framework (Von Neumann & Morgenstern, 1944). Conversely, when nature's strategy is unknown, the player faces a problem of *decision making under uncertainty*, a setting that requires a more axiomatic treatment (Milnor, 1951; Savage, 1951; Papadimitriou, 1985). We defer to Appendix A for an extended discussion of this topic.

Common Agency. In many scenarios, multiple parties (or principals) have preferences over the actions an agent performs. We refer to those cases where an agent must satisfy the preferences of multiple principals, which may conflict as *common agency problems* (Bernheim & Whinston, 1986; Peters, 2001). This formulation assumes that principals have preferences over every pair of actions $y, y' \in \mathcal{Y}$ in the agent action space, and that those preferences are expressed through a set of incentives or payoff functions $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_M\}$, one for each of the M principals, such that each is a function $\ell_k : \mathcal{Y} \to \mathrm{R}$ that scores the agent's actions.

Machine Learning. If one restates SL and RL on interactive terms (see Appendix B), their definitions reveal functionally equivalent components. Conveniently, *instances* and *states* (x_t and s_t), *outputs* and *actions* (y_t and a_t), *data distributions* and *transition functions* (μ and ρ), models and policies (μ and π), or *reward functions* and *metric functions* (ℓ and r) have equivalent roles. Meanwhile, the distinctions may only be relevant for designing learning algorithms. Instead, we focus on common components, the triplet $\langle \pi, \rho, \ell \rangle$, and their implied multiplayer structure.

3 Milnor-Myerson Games

The triplet of components $\langle \pi, \rho, \ell \rangle$ describes a multiplayer structure that connects ML to decision theory and principal-agent problems.

The Milnor Decomposition. In decision theory, when the decision-making player is faced with a player whose acts may affect the decision-maker payoffs but has no interest in the game outcome of the interactions, the decision-making player is playing a *game against nature* (Milnor, 1951). In

[2]In honour to John Milnor's and Robert B. Myerson's pioneering contributions to decision theory (Milnor, 1951) and principal-agent problems (Myerson, 1983).

either SL or RL problems, the data distribution or the transition dynamics describe those world-grounded processes. Nature's behaviour is obtained by either human interactions with the (true) data distribution and recollected on datasets (sketches) (LeCun et al., 2010; Deng et al., 2009; Lin et al., 2014) or simulated to leave agents to interact with it (Brockman et al., 2016).

Proposition 1. *Every machine learning problem models a multi-player game against nature.*

Proof. Sketch: Start from a *game against nature* and consider a sequential, repeated, perfect information game where nature has a known fixed strategy and the agent maximizes expected payoffs. □

ML makes several assumptions about nature's action space structure to model world-grounded problems. Thus, the action space $\mathcal{X}$ may represent images (LeCun et al., 2010; Russakovsky et al., 2014), the gameboards of *chess*, *shogi* or *go* (Silver et al., 2016; 2018), vector-based observations for robotic control (Brockman et al., 2016; Andrychowicz et al., 2018), or language tokens (Devlin et al., 2018; Radford et al., 2019). In (Markovian) RL, the *Markov* assumption constrains nature's actions to be sufficient statistics of past interactions (Puterman, 2005). Moreover, SL and RL differ in their assumption about nature's strategy. In the first, the data distribution is a *memoryless* (or reactive) strategy $\rho : \Delta(\mathcal{X})$, and as such, nature's actions are *independent* from previous actions[3]. In the latter, the Markovian transition dynamics represent a memory-one strategy $\rho : \mathcal{X} \times \mathcal{Y} \to \Delta(\mathcal{X})$ such that the last interaction conditions every action.

Myerson Decomposition. The traditional interpretation of agency in the context of ML problems, mainly in RL problems, is one where an agent is an entity that changes the state of the environment through its actions to achieve some goal (Sutton & Barto, 2018). This *causal* interpretation often neglects the *representative* perspective of agency, common in economics (Ross, 1973), entertainment (Zelenski, 2003), or sports (Shropshire & Davis, 2008), where an agent is an entity that acts on behalf of other(s). While in the real world, both perspectives coexist and are the reason that *principal-agent* problems arise (Ross, 1973; Mitnick, 1975; Myerson, 1983), we believe that the *agent-as-a-representative* perspective uncovers novel aspects of artificial agents and ML problems, as we will show in Sec. 4. For instance, an artificial agent-principal relationship avoids some traditional principal-agent problems.

Remark 1. *Artificial agents have no intrinsic objectives and only maximize those (extrinsically) specified by their designers (Nicholson, 2013; Roli et al., 2022; Jaeger, 2023; Jaeger et al., 2024).*

Thus, our work takes a philosophical stance in the *extrinsic* vs *intrinsic* objective debate, common in RL neuro-cognitive foundations (Singh et al., 2009; Hassabis et al., 2017; Lake et al., 2017; Botvinick et al., 2020).

those preferences appear in ML in two forms. In SL, the accuracy function generally has the form $\ell(x, y) = 1$ if $y = y^*$ where a unique agent action $y^* \in \mathcal{Y}$ (e.g., ground-truth label) is the best response to nature's action $x \in \mathcal{X}$. The accuracy function expresses a top-rank choice preference $y^* \succ y'$ for such action, and the principal's indifference $y \sim y'$ for other actions $y, y' \in \mathcal{Y}$. In RL, a (Markov) reward function $\ell(x, y)$ expresses a (possible partial) set of preferences over agent actions $y \succeq_x y'$, conditioned on nature's action $x \in \mathcal{X}$.

Higher-Order Milnor-Myerson Games. Beyond the *canonical* structure containing an agent, nature, and a principal, denoted by $\text{SASP}_\text{N} = \langle \pi, \rho, \ell \rangle$, there are also Milnor-Myerson games of a higher order. In the Multi-Agent Single Principal (MASP_N) game, that we denote by $\text{MASP}_\text{N} = \langle \Pi, \rho, \ell \rangle$, a set of agents $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$ play against nature as representatives of a common principal whose strategy $\ell : \mathcal{X} \times \boldsymbol{\mathcal{Y}} \to \mathrm{R}^n$ decides every player's payoff, and $\boldsymbol{\mathcal{Y}} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots \times \mathcal{Y}_n$ denotes the n-player joint action space. If we consider a common agency problem, a Single-Agent Multi-Principal (SAMP_N) game, denoted by the tuple $\text{SAMP}_\text{N} = \langle \pi, \rho, \mathcal{L} \rangle$, models a structure where an agent π, plays against nature's strategy ρ, on behalf of a set of principals $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_m\}$, whose strategies $\ell_k : \mathcal{X} \times \mathcal{Y} \to \mathrm{R}$ define the agent's payoff. Then, in the Multi-Agent and Multi-Principal (*MAMPn*) structure, that we denote by $\text{MAMP}_\text{N} = \langle \Pi, \rho, \mathcal{L} \rangle$, a

[3]This is an alternative interpretation of the *independence* component of the *i.i.d* assumptions (Bishop, 2006).

set of agents $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$, play a game against nature's strategy ρ on behalf of multiple principals $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_m\}$, whose strategies $\ell_k : \mathcal{X} \times \boldsymbol{\mathcal{Y}} \to \mathrm{R}^n$ define the m payoffs every player receives. These higher-order structures model interactions in several ML problems (Appendix C.1)

4 The Artificial Principal-Agent Problem

Understanding the foundation that ML problems share through Milnor-Myerson games and their higher-order structures presents novel challenges (see **??**). From all, we focus on the human designer-artificial agent relationship. We make precise the modes of principal supervision and how they connect to sparse and dense objectives. Next, we describe how ML algorithms propagate the principal's preferences to later, in Sec. 5, leverage these ideas to connect *AlphaGo* and LLMs.

4.1 Modes of Principal Supervision

In traditional principal-agent problems, principals incur agency costs. These costs include the costs of monitoring, specification, and others (Jensen & Meckling, 1976). In artificial agency, the *costs of specification* limit the principal's strategy effective domain. In many ML problems, specifying payoffs for every $x \in \mathcal{X}$ is costly or unfeasible for large (or infinite) action spaces. Generally, principals' strategies are defined over a subset $\mathcal{X}' \subseteq \mathcal{X}$ of nature's action space. The notions of supervision that follow are limited to such a subset.

Weak Supervision. We defined *sparse* payoff functions, with domain on $\mathcal{X}'$, to be those principal strategies that do not communicate preferences over certain outcomes. More formally,

Definition 1. *Let $\ell : \mathcal{X}' \times \mathcal{Y} \to \mathrm{R}$ denote the principal's strategy. The strategy is **weakly sparse** if there exists a nature's action $x \in \mathcal{X}'$ where it induces no order over the player's action set $\mathcal{Y}(x)$. For those nature's actions, the principal expresses an indifference $y \sim_x y'$ for every action pair $y, y' \in \mathcal{Y}(x)$ the payoffs are equal and constant $\ell(x, y) = \ell(x, y')$.*

A stronger and more useful notion of sparseness is one where, except on some *terminal* interactions where there should be at least one preferred outcome, the principal only expresses indifference in most states of nature.

Definition 2. *Let $\ell : \mathcal{X}' \times \mathcal{Y} \to \mathrm{R}$ denote the principal's strategy. The strategy is **strongly sparse** if for every nature's action $x \in \mathcal{X}'$ it induces no order over the player's action set $\mathcal{Y}(x)$. The principal expresses indifference $y \sim_x y'$ for every pair of actions $y, y' \in \mathcal{Y}(x)$ whose payoffs are equal and constant $\ell(x, y) = \ell(x, y')$.*

Example 1. *Among the strongly sparse principal strategies, we have those leveraged to obtain the gameplaying abilities AlphaGo, AlphaGo Zero, and AlphaZero (Silver et al., 2016; 2017; 2018).*

Weakly and strongly sparse payoffs are a form of weak principal supervision or indifference.

Strong Supervision. In dense payoff functions, the principal explicitly communicates her preferences over actions. In the first of such cases, at least one agent action is preferred over other actions as a response to nature's actions or states. More formally,

Definition 3. *Let $\ell : \mathcal{X}' \times \mathcal{Y} \to \mathrm{R}$ denote a principal's strategy. The strategy is **weakly dense** if, for every nature action $x \in \mathcal{X}'$, it induces a partial order over the player's action set $\mathcal{Y}(x)$. The principal expresses a strong preference for outcome $y^* \succ_x y$ by having $\ell(x, y^*) > \ell(x, y')$ on every action pairs $y^*, y \in \mathcal{Y}$.*

Example 2. *Most (multiclass) supervised classification problems (Bishop, 2006) are weakly dense, where one* label *$y \in \mathcal{Y}$ is preferred over others given a certain instance $x \in \mathcal{X}$.*

Definition 4. *Let $\ell : \mathcal{X}' \times \mathcal{Y} \to \mathrm{R}$ denote a principal's strategy. The strategy is **strongly dense** if for every nature action $x \in \mathcal{X}'$, it induces a total order over the player action set $\mathcal{Y}(x)$. The principal expresses complete and transitive preferences $y \succ_x y'$ or $y' \succ_x y$ for every pair of actions $y, y' \in \mathcal{Y}(x)$ such that either $\ell(x, y) > \ell(x, y')$ or $\ell(x, y') > \ell(x, y)$.*

Example 3. *In RL, shaped rewards for goal-based RL problems (Ng et al., 1999; Gupta et al., 2022) represent instances of strongly dense principal strategies.*

We consider both weakly and strongly dense payoff functions a form of *strong principal supervision*, as they encapsulate the principal's explicit preferences over outcomes.

4.2 Algorithms as Preferences Propagation Mechanisms

In RL and SL problems, principal preferences induced by its strategy propagate forward through the agent-nature sequential interactions. Thus, learning algorithms produce *optimal strategies* as objects that hold the consequences of maximizing the principal's preferences over time.

Reinforcement Learning. In RL, algorithms that leverage *state-action-value* functions (Watkins, 1989; Mnih et al., 2015), push forward the preferences encoded by the reward function $\ell(x, y) \to Q^*(x, y)$ towards the optimal action-value function $Q^*(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathrm{R}$. This function holds the maximum expected sum of payoffs the agent receives from the principal over $t \in \mathrm{N}$ interactions with nature's memory-one strategy $\rho : \mathcal{X} \times \mathcal{Y} \to \Delta(\mathcal{X})$, if takes action $y \in \mathcal{Y}$, in nature state $x \in \mathcal{X}$ (Sutton & Barto, 2018). Thus, it holds preferences over actions $y \succ_x^t y'$, that can be derived by comparing values $Q^*(x, y) > Q^*(x, y')$. Similarly, *policy gradient* (Sutton et al., 1999) and *actor-critic* algorithms (Konda & Tsitsiklis, 1999; Schulman et al., 2017; Haarnoja et al., 2018), push the preferences in the reward function $\ell(x, y) \to \pi^*(y|x)$ towards the optimal policy $\pi^*(y|x)$ (i.e., the actor). The probabilities $\pi^*(y|x) > \pi^*(y'|x)$ encode the t-step preferences $y \succ_x^t y'$ over actions.

Supervised Learning. In SL, nature's actions are independent of the agent's. Therefore, the principal's preferences do not propagate through interactions. Instead, learning algorithms transform the principal preferences $\ell(x, y) \to \pi^*(y|x)$ directly into an *optimal model* $\pi^*(y|x)$ by interpreting the principal's *0-1* strategy through probabilistic lenses (e.g., the principle of maximum likelihood Bishop (2006)). Thus, the optimal strategy $\pi^*(y|x)$ encodes the probability that action $y \in \mathcal{Y}$ is the top-ranked action (i.e., $y \succ y'$ for all $y' \in \mathcal{Y}$) as a response to nature action $x \in \mathcal{X}$. In the presence of multiple principals, like in the LLM example above, a single top-ranked action may not exist. Multi-label classification or label ranking solve similar problems (Tsoumakas & Katakis, 2007; Vembu & Gärtner, 2011).

Stationarity Under Propagation. We argue that principal strategies whose expressed preferences are unchanged by algorithms' forward propagation mechanisms facilitate learning.

Definition 5. *Let $\ell : \mathcal{X}' \times \mathcal{Y} \to \mathrm{R}$ be the principal strategy in $\mathcal{G} = \langle \rho, \pi, \ell \rangle$, the **principal preferences are stationary** if for every nature action $x \in \mathcal{X}$, whenever $y \succ_x y'$, for every pair of actions $y, y' \in \mathcal{Y}(x)$, the pushed-forward preferences $y \succ_x^t y'$ remain unchanged.*

There are many instances in RL where the principal strategies that may hold this stationarity property have accelerated agent learning. For instance, shaped rewards may belong to this class of strategies (Ng et al., 1999; Gupta et al., 2022). Likewise, rewards learned from demonstrations using inverse reinforcement learning may behave similarly (Ng & Russell, 2000; Ziebart et al., 2008).

5 AlphaGo, LLMs, and The Principle of Indifference

5.1 From AlphaGo to AlphaZero

The problems of learning to play *go*, *chess*, and *shogi* share the structure of an MASPN game with two players and a principal, that we denoted by $\mathcal{G}_1^2 = \langle \rho, \Pi, \ell \rangle$. In these games, the two players, with action spaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$, and strategies $\Pi = \{\pi_1, \pi_2\}$, repeatedly play *against* nature, whose action space $\mathcal{X}$ models the state of the gameboard, and responds to players moves with a (deterministic) strategy $\rho : \mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2 \to \mathcal{X}$. A fundamental challenge to applying ML to these board games is the lack of evaluation criteria. Formulating a principal strategy $\ell : \mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2 \to \mathrm{R}$ that determines how valuable a position is seems unfeasible (Ramon et al., 2001; Müller, 2002; Gelly & Silver, 2008; McGrath et al., 2022). Instead, the principal's strategy $\ell : (\mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2)^n \to \mathrm{R}$ is defined such that

after a finite sequence of interactions $n \in \mathrm{N}$, players either *win*, *lose*, or *draw* the game, according to some termination conditions and receive final payoffs (Silver et al., 2017; 2018), an instance of principal weak supervision.

5.2 Large Language Models

LLMs pre-training can be described by a SAMPn game, that we denote by $\mathcal{G}_m^1 = \langle \rho, \pi, \mathcal{L} \rangle$, where the LLM is a common agent that satisfies the preferences of m principals (for very large values of $m \in \mathrm{N}$) with strategies $\ell_k \in \mathcal{L}$. In this setting, an LLM plays with strategy $\pi(y|\boldsymbol{x})$ where $\boldsymbol{x} \in \mathcal{X}^n$ denotes an input sequence of $n \in \mathrm{N}$ words or tokens, produced by nature's strategy $\rho : \mathcal{X}^n \times \mathcal{Y} \to \mathcal{X}^{n+1}$, and $y \in \mathcal{Y}$ is a possible next-token completion. Then, in Internet-scale datasets leveraged to train LLM, every entity that uttered these sentences turns into an *implicit principal* of the model. Their strategies are expressed by an accuracy function $\ell_k : \mathcal{X}^n \times \mathcal{Y} \to \{0, 1\}$ (or its differentiable surrogates (Bishop, 2006)). The principal's ignorance problem emerges by aggregating every principal's preferences. For instance, suppose that for the sequence or prompt $\boldsymbol{x}$ = *"The sky is"*, values of $y = \{$*blue*, *gray*, *beautiful*$\}$ are plausible model completions. Then, sentences like *"The sky is blue"* or *"The sky is beautiful"* in a dataset represent principals' *incompatible* preferences (e.g., *blue* $\succ$ *gray* $\sim$ *beautiful* $\sim \ldots$ and *beautiful* $\succ$ *gray* $\sim$ *blue* $\sim \ldots$) over the space of completions of that phrase, that may create an aggregated partial preference relationship *blue* $\sim_{\boldsymbol{x}}$ *beautiful* $\succ_{\boldsymbol{x}}$ *gray*$\ldots$ over the possible completions.

5.3 The Principle of Indifference

It is in the principal's best interest to define stationary strategies. One may immediately wonder why a principal with foresight over her preferences would require a learning algorithm. We argue that the strategy specification (i.e., for the learning problem) may be simpler than writing the program that executes them. And that trade-off is what favours *learning* over, for instance, *knowledge-based systems*. However, specifying them would require her to have the computational capacity to foresee the long-term consequences of the stepwise preferences. There is a space of *small world* problems (Binmore, 2007; 2017) where the principal may be able to provide such *strong supervision*, but in other *large world* problems, those one may consider attractive, rarely human principals have that foresight because of their *boundedness* (Simon, 1990; Selten, 1990; Tversky & Kahneman, 1974), and thus, are generally unable to perform such computations. Consequently, rationally-bounded principals may introduce biased preferences through their strategies. Both *AlphaGo* and LLMs represent cautionary tales on this issue. For instance, the initial *AlphaGo* playing strategy was bootstrapped from expert human demonstrations (Silver et al., 2016). Yet, *AlphaGo Zero* and *AlphaZero* discarded those demonstrations, learned the games *tabula rasa* from strongly sparse supervision, and achieved far superior gameplaying abilities (Silver et al., 2017; 2018). Similarly, LLMs are pre-trained, as we explained earlier, with a multi-principal structure that induces indifference over preferences, but fine-tuning (Ouyang et al., 2022), a stronger form of supervision, hinders capabilities built by pre-training (Wang et al., 2024).

Consequently, the *weak* vs *strong* supervision dichotomy presented in Sec. 4 has implications beyond traditional concerns about the hardness of the credit assignment problem (Minsky, 1961; Sutton, 1984). Strong supervision imposes some computational and rationality requirements on human designers that rarely hold and, as such, almost guarantee biased solutions. The general surprise around *AlphaGo*'s Move 37 provides a good context to this observation. Both *AlphaGo* or *AlphaZero* and LLMs benefited from weak supervision in the form of principal indifference. We argue *indifference* should be a guiding principle to build objectives for artificial agents that can think and act as humans but also be able to innovate (McGrath et al., 2022).

Definition 6. ***(The Principal's Principle of Indifference)*** *If the principal is uncertain about the consequences of her step-wise preference among agent actions, her strategy should express indifference to those actions.*

Undoubtedly, the principle of indifference offers principals no free lunch. Beyond the credit assignment problem, weak supervision leads to loss of designer control and revitalizes the problems of AI safety (Amodei et al., 2016). Moreover, it requires principals to *metareason* about their bounded rationality and its influence on their understanding of the problem space.

6 Machine Learning Through Milnor-Myerson Games

Multiplayer is The Question and The Answer. Understanding ML problems under the Milnor-Myerson games has many implications. Many have argued that multiplayer interactions are fundamental to obtaining extremely capable agents (Shoham et al., 2007; Stone, 2007; Leibo et al., 2019; Baker et al., 2020). The ideas we presented here may further reinforce the beliefs that most outstanding ML results originate from multiplayer interactions at scale.

Axiomatic Decision Theory and ML. Interestingly, the objectives in Eq. 1 and Eq. 2 are formulated, from a decision theory perspective, as problems of *decision-making under risk*, where nature's strategy is assumed known and the agent maximizes *expected* payoff (Peterson, 2017). Except for a very limited set of real-world problems (e.g., *chess*, *go*, videogames, or simple robotics), nature strategy is hardly computable. It is hard to describe such a function at the scale of recent LLMs. Thus, we may need to reconcile the objectives that describe ML paradigms with critiques to (Bayesian) decision theory on large-scale problems (Savage, 1954; Binmore, 2007; 2017).

The Principal-Artificial Agent Problem. In economics and mechanism design, in the *principal-agent problem* an *agent* may have objectives that are misaligned with the *principal*'s (Ross, 1973; Myerson, 1983; Conitzer & Sandholm, 2002). While, in theory, learning algorithms would continuously optimize their designer's objective (Silver et al., 2021), in practice, several other problems arise that cause the artificial agent and its designer(s) to have misaligned objectives. Specification problems lead to learning agents displaying behaviours that diverge from the designer's intended goals Amodei et al. (2016); Leike et al. (2018). In RL, there is still a debate on whether reward functions are sufficiently expressive instruments of the designer's goals (Silver et al., 2021) or if they impose some expressivity constraints (Abel et al., 2021; Vamplew et al., 2022; Bowling et al., 2023). In ML, Hadfield-Menell & Hadfield (2018) introduces the principal-agent problem through contract theory. Similarly, LaCroix & Bengio (2019) defines equivalences between economic and ML concepts, including the principal-agent problem. Both works present the principal-agent problem and bounded rationality in AI safety perspective (Amodei et al., 2016). In contrast, we consider AI alignment to be one of the several consequences of the principal-agent problem. In the opposite direction, from ML to other areas, Ben-Porat et al. (2023) leveraged ML techniques, particularly RL reward-shaping techniques, to improve the principal's outcome in a traditional principal-agent problem, and Gan et al. (2023) study the principal-agent problem under partial observability and communication constraints. Hyland et al. (2023) specification and verification in principal-agent problems by distributed computation of boolean games (Harrenstein et al., 2001).

The Limits of Our Understanding. In our framing, understanding these components as *nature* raises an important issue. If the principals, due to their cognitive limitations, have a limited understanding of the world, artificial agents' behaviour would be limited by the principal's ability to accurately devise sketches of nature's strategy (Sadeghi & Levine, 2016; Chebotar et al., 2018) and, ultimately, to produce novel problems (Leibo et al., 2019). These two situations further reinforce the *principal-agent* problem between designers and their artificial learning agents.

The In-Roads to Unshackled Artificial Agents. The question is whether it is possible to release artificial agents from their *representatives* duties. Several existing threads may lead to generally capable agents. For instance, open-ended learning (Wang et al., 2019; Stooke et al., 2021; Sigaud et al., 2023; Abel et al., 2023; Bruce et al., 2024) or intrinsic motivation objectives (Schmidhuber, 1991; Oudeyer & Kaplan, 2009) offer promising results. Furthermore, approaches to *Embodied AI* envision artificial agents that interact directly with their environments (Puig et al., 2023). However, unshackled agents may develop preferences that are not aligned with their human designers, and then a true principal-agent problem will materialize.

References

David Abel, Will Dabney, A Harutyunyan, Mark K Ho, M Littman, Doina Precup, and Satinder Singh. On the expressivity of markov reward. *Advances in neural information processing systems*, abs/2111.00876, November 2021.

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. July 2023.

Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024. URL https://www.marl-book.com.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, pp. 29, 2016.

Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous In-Hand manipulation. *arXiv [cs.LG]*, August 2018.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from Multi-Agent autocurricula. In *International Conference on Learning Representations*, 2020.

Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, 6(5): 679–684, 1957.

Omer Ben-Porat, Yishay Mansour, Michal Moshkovitz, and Boaz Taitler. Principal-Agent reward shaping in MDPs. December 2023.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Jozefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P d Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *arxiv*, December 2019.

B Douglas Bernheim and Michael Whinston. Common agency. *Econometrica: journal of the Econometric Society*, 54(4):923–942, 1986.

Ken Binmore. Rational decisions in large worlds. *Annals of economics and statistics*, (86):25–41, 2007.

Ken Binmore. On the foundations of decision theory. *Homo Oeconomicus*, 34(4):259–273, December 2017.

Christopher M Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

Matthew Botvinick, Jane X Wang, Will Dabney, Kevin J Miller, and Zeb Kurth-Nelson. Deep reinforcement learning and its neuroscientific implications. Technical report, 2020.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3003–3020. PMLR, 2023.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. June 2016.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. February 2024.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. March 2023.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the Sim-to-Real loop: Adapting simulation randomization with real world experience. October 2018.

Vincent Conitzer and Thomas Sandholm. Complexity of mechanism design. In *UAI*, 2002.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, June 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. October 2018.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient scaling of language models with Mixture-of-Experts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5547–5569. PMLR, 2022.

R Duncan Luce and Howard Raiffa. *Games and Decisions: Introduction and Critical Survey*. Courier Corporation, April 1989.

M Everingham, L Van Gool, C K I Williams, J Winn, and A Zisserman. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, June 2010.

Jiarui Gan, Rupak Majumdar, Debmalya Mandal, and Goran Radanovic. Sequential Principal-Agent problems with communication: Efficient computation and learning. June 2023.

Sylvain Gelly and David Silver. Achieving master level play in 9×9 computer go. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 3*, AAAI'08, pp. 1537–1540. AAAI Press, July 2008.

Gemini Team. Gemini: A family of highly capable multimodal models. December 2023.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, S Kakade, and S Levine. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. *Advances in neural information processing systems*, abs/2210.09579, October 2022.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic algorithms and applications. December 2018.

Dylan Hadfield-Menell and Gillian Hadfield. Incomplete contracting and AI alignment. April 2018.

Paul Harrenstein, Wiebe van der Hoek, John-Jules Meyer, and Cees Witteveen. Boolean games. In *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '01, pp. 287–298, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607919.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-Inspired artificial intelligence. *Neuron*, 95(2):245–258, July 2017.

D Hyland, Julian Gutierrez, and M Wooldridge. Principal-Agent boolean games. *International Joint Conference on Artificial Intelligence*, pp. 144–152, May 2023.

Johannes Jaeger. Artificial intelligence is algorithmic mimicry: why artificial "agents" are not (and won't be) proper agents. June 2023.

Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh. Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in psychology*, 15, June 2024.

Michael C Jensen and William H Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3(4):305–360, October 1976.

Greg Kohs. AlphaGo: The movie, 2017.

Vijay Konda and John Tsitsiklis. Actor-Critic algorithms. In S Solla, T Leen, and K Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, April 2009.

Travis LaCroix and Yoshua Bengio. Learning from learning machines: Optimisation, rules, and social norms. December 2019.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *The Behavioral and brain sciences*, 40, 2017.

Yann LeCun, Corinna Cortes, and Chris Burges. MNIST handwritten digit database, 2010.

Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv:1903.00742 [cs, q-bio]*, March 2019.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. November 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. May 2014.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 157–163. Elsevier, 1994.

J McCarthy, M L Minsky, N Rochester, and C Shannon. A proposal for the darmouth summer research project on artificial intelligence. Technical report, August 1955.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in AlphaZero. *Proceedings of the National Academy of Sciences of the United States of America*, 119 (47):e2206625119, November 2022.

John Menick. Move 37: Artificial intelligence, randomness, and creativity. *Mousse Magazine*, 55 (53), 2016.

John Milnor. Games against nature. Technical report, RAND PROJECT AIR FORCE SANTA MONICA CA, 1951.

Marvin Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961. doi: 10.1109/JRPROC.1961.287775.

Barry M Mitnick. The theory of agency: A framework. *Barry M. Mitnick, The Theory of Agency (Cambridge University Press, Forthcoming)*, 1975.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, February 2015.

Martin Müller. Counting the score : Position evaluation in computer go. Technical report, 2002.

Roger B Myerson. Mechanism design by an informed principal. *Econometrica: journal of the Econometric Society*, 51(6):1767–1797, 1983.

Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 663–670, 2000.

Andrew Y Ng, Daishi Harada, and Stuart J Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 278–287, San Francisco, CA, USA, June 1999. Morgan Kaufmann Publishers Inc.

Daniel J Nicholson. Organisms≠Machines. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 44(4, Part B):669–678, December 2013.

OpenAI. GPT-4 technical report. March 2023.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Christos H Papadimitriou. Games against nature. *Journal of Computer and System Sciences*, 31 (2):288–301, October 1985.

Michael Peters. Common agency and the revelation principle. *Econometrica: journal of the Econometric Society*, 69(5):1349–1372, 2001.

Martin Peterson. *An Introduction to Decision Theory*. Cambridge University Press, March 2017.

Silviu Pitis. Rethinking the discount factor in reinforcement learning: a decision theoretic approach. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, number Article 975 in AAAI'19/IAAI'19/EAAI'19, pp. 7949–7956. AAAI Press, January 2019.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A Co-Habitat for humans, avatars and robots. October 2023.

Martin L Puterman. *Markov Decision Processes*. 2005.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Roxana Rădulescu, Patrick Mannion, Diederik M Roijers, and Ann Nowé. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous agents and multi-agent systems*, 34(1), April 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. May 2023.

Jan Ramon, Tom Francis, and Hendrik Blockeel. Learning a go heuristic with tilde. In *Computers and Games*, pp. 151–169. Springer Berlin Heidelberg, 2001.

D Roijers, P Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *The journal of artificial intelligence research*, abs/1402.0590, October 2013.

Andrea Roli, Johannes Jaeger, and Stuart A Kauffman. How organisms come to know the world: Fundamental limits on artificial general intelligence. *Frontiers in Ecology and Evolution*, 9, 2022.

Willem Röpke, Carla Groenland, Roxana Rădulescu, Ann Nowé, and Diederik M Roijers. Bridging the gap between single and multi objective games. January 2023.

Stephen A Ross. The economic theory of agency: The principal's problem. *The American economic review*, 63(2):134–139, 1973.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. September 2014.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2020.

Fereshteh Sadeghi and Sergey Levine. CAD2RL: Real Single-Image flight without a single real image. November 2016.

L J Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46 (253):55–67, 1951.

Leonard Jimmie Savage. *The Foundations of Statistics*. Wiley, 1954.

J Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats*. The MIT Press, 1991.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. July 2017.

Reinhard Selten. Bounded rationality. *Journal of institutional and theoretical economics: JITE = Zeitschrift fur die gesamte Staatswissenschaft*, 146(4):649–658, 1990.

L S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial intelligence*, 171(7):365–377, May 2007.

K L Shropshire and T Davis. *The Business of Sports Agents*. University of Pennsylvania Press, Incorporated, 2008.

Olivier Sigaud, Gianluca Baldassarre, Cedric Colas, Stephane Doncieux, Richard Duro, Nicolas Perrin-Gilbert, and Vieri Giuliano Santucci. A definition of Open-Ended learning problems for Goal-Conditioned agents. November 2023.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, October 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, December 2018.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial intelligence*, pp. 103535, May 2021.

Herbert A Simon. Bounded rationality. In John Eatwell, Murray Milgate, and Peter Newman (eds.), *Utility and Probability*, pp. 15–18. Palgrave Macmillan UK, London, 1990.

Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from? https://all.cs.umass.edu/pubs/2009/singh_l_b_09.pdf, 2009. Accessed: 2023-2-15.

Peter Stone. Multiagent learning is not the answer. it is the question. *Artificial intelligence*, 171(7): 402–405, May 2007.

Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard, and Wojciech Marian Czarnecki. Open-Ended learning leads to generally capable agents. July 2021.

R Sutton, David A McAllester, Satinder Singh, and Y Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, pp. 1057–1063, November 1999.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2nd edition, 2018.

Richard Stuart Sutton. *TEMPORAL CREDIT ASSIGNMENT IN REINFORCEMENT LEARNING*. PhD thesis, Ann Arbor, United States, 1984.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

A M Turing. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind; a quarterly review of psychology and philosophy*, LIX(236):433–460, October 1950.

A Tversky and D Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185 (4157):1124–1131, September 1974.

Peter Vamplew, Benjamin J Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M Roijers, Conor F Hayes, Fredrik Heintz, Patrick Mannion, Pieter J K Libin, Richard Dazeley, and Cameron Foale. Scalar reward is not enough: a response to silver, singh, precup and sutton (2021). *Autonomous agents and multi-agent systems*, 36(2):41, July 2022.

Shankar Vembu and Thomas Gärtner. Label ranking algorithms: A survey. In Johannes Fürnkranz and Eyke Hüllermeier (eds.), *Preference Learning*, pp. 45–64. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P Agapiou, Max Jaderberg, Alexander S Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, November 2019.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, Princeton, NJ, 1944.

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. Paired Open-Ended trailblazer (POET): Endlessly generating increasingly complex and diverse learning environments and their solutions. January 2019.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. Two-stage LLM fine-tuning with less specialization and more generalization. In *The Twelfth International Conference on Learning Representations*, 2024.

Christopher Watkins. *Learning from delayed rewards*. PhD thesis, 1989.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, August 2012.

David Zelenski. Talent agents, personal managers, and their conflicts in the new hollywood. *Southern California law review*, 76(4), May 2003.

Cha Zhang and Yunqian Ma (eds.). *Ensemble Machine Learning: Methods and Applications*. Springer New York, 2012.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, pp. 1433–1438. AAAI Press, 2008. ISBN 9781577353683.

A Games Against Nature

To better comprehend what games against nature represent, we reproduce below the following dilemma that we will call *Savage's Sixth Egg Dilemma* (Savage, 1954), adapted from (Duncan Luce & Raiffa, 1989), Chapter 13.

"Your spouse has just broken five good eggs into a bowl when you come in and volunteer to finish making the omelet. A sixth egg, which for some reason must be either used or wasted altogether, lies unbroken beside the bowl. You must decide what to do with this unbroken egg. Perhaps it is not too great an oversimplification to say that you must decide among three acts only, namely, to break it into a bowl containing the other five, to break it into a saucer for inspection, or to throw it away without inspection. Depending of the state of the egg, each of those three acts will have some consequence of concern to you, say as indicated in the following table:"

Action	**State of Nature**	
	Good	**Rotten**
Break into bowl	Six-egg omelet	No omelet, five good eggs trashed
Break into saucer	Six-egg omelet, saucer to wash	Five-egg omelet and a saucer to wash
Throw away	Five-egg omelet, good egg destroyed	Five-egg omelet

Table 1: Savage's Sixth Egg Dilemma.

For instance, the principles of *insufficient reason* (Milnor, 1951), *minimax* (Savage, 1954) or *minimax regret* (Savage, 1951) present some axiomatic approaches a rational agent could follow to inform its decisions when faced with an uncertain prospect over nature's actions, meaning, the agent does not know $\rho \in \Delta(\mathcal{X})$. However, the more predominant formulation of sequential decision-making problems is linked to the principle of *expected utility* (Von Neumann & Morgenstern, 1944). A derivative of the problem of *decision making under risk*, the expected utility theory approaches *decision making* assuming the decision-making agent knows the probabilities $\rho \in \Delta(\mathcal{X})$ of *nature*'s actions (Duncan Luce & Raiffa, 1989; Peterson, 2017).

B Interactive Machine Learning Problems

We re-state SL and RL on interactive terms to understand how our discussion applies to both.

Supervised Learning. A supervised learning task (Bishop, 2006), denoted by the tuple $\mathrm{T} = \langle \mathcal{X}, \mathcal{Y}, \rho, \ell \rangle$, is a problem where instances $x \in \mathcal{X}$ are drawn from a data distribution $\rho \in \Delta(\mathcal{X})$, and presented to a *model* $\mu : \mathcal{X} \to \Delta(\mathcal{Y})$ to produce outputs $y \in \mathcal{Y}$ scored by a *metric function* $\ell : \mathcal{X} \times \mathcal{Y} \to \mathrm{R}$. The objective of supervised learning is to find a model $\mu^* \in \mathrm{M}$ that maximizes the cumulative performance function $\mathcal{J}$ for a number $T \in \mathbb{N}$ of interactions between the model and the data distribution, scored by the metric function.

$$\mathcal{J}(\mu; \rho, \ell) = \mathbb{E}_{\substack{x_t \sim \rho \\ y_t \sim \mu(\cdot|x_t)}} \left[\sum_{t=1}^{T} \ell(x_t, y_t) \right] \tag{1}$$

Notice that our presentation offers a more interactive but equally valid perspective of SL. Generally, datasets contain pre-populated instance-outputs pairs, representing a *history of past interactions* with the data distribution, while also holding the evaluation values of the metric function[4].

Reinforcement Learning. In a traditional RL problem (Sutton & Barto, 2018), the tuple $\mathrm{T} = \langle \mathcal{S}, \mathcal{A}, \rho, r, \gamma \rangle$ and defines a Markov Decision Process (MDP) (Bellman, 1957; Puterman, 2005), a

[4]This perspective is also adopted by other approaches like offline RL. See (**?**).

sequence of *states* $s \in \mathcal{S}$ are drawn from a transition function $\rho : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and presented to a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ to produce actions $a \in \mathcal{A}$, scored by a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathrm{R}$. The objective in RL is to find a policy $\pi^* \in \Pi$ that maximizes a the performance $\mathcal{J}$ function:

$$J(\pi; \rho, r, \gamma) = \mathbb{E}_{\substack{s_t \sim \rho(\cdot|s_{t-1}, a_{t-1}) \\ a_t \sim \pi(\cdot|s_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \tag{2}$$

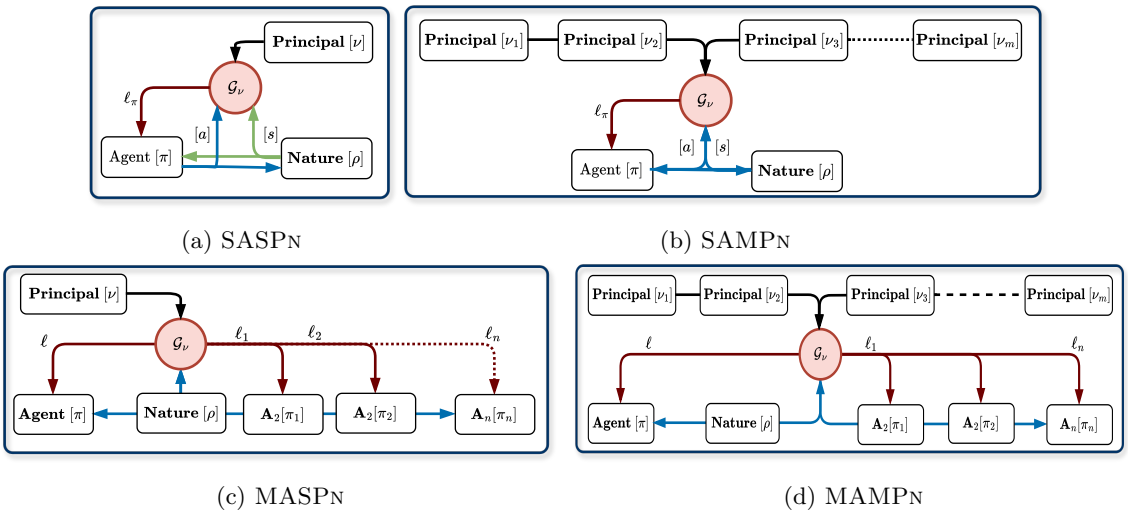
where the interactions between the policy and the transition dynamics, scored by the reward function, and the discount factor $\gamma \in (0, 1]$ introduces a preference for the value of earlier interactions (Pitis, 2019).

Learning Algorithms. In a broad sense, we consider a *learning algorithm* to be any procedure to search and find a solution, or at least a close approximation, for the objective:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathcal{J}(\pi, \rho, r) \tag{3}$$

such that $\pi^* \in \Pi$ satisfies $\mathcal{J}(\pi^*) \geq \mathcal{J}(\pi)$, and the evaluation functional $\mathcal{J}$ originates in Eq. 1 or 2.

C A Graphical Representation of Milnor-Myerson Games



(a) SASPN (b) SAMPN

(c) MASPN (d) MAMPN

Figure 1: A graphical representation of Milnor-Myerson games and the higher order structures.

C.1 Milnor-Myerson Games and Where to Find Them

The MASPN is found in MARL (Shapley, 1953; Littman, 1994; Albrecht et al., 2024), and, in SL, on *ensemble models* (Zhang & Ma, 2012) and *mixture of experts* (Yuksel et al., 2012), two techniques recently employed in LLM (Du et al., 2022). The SAMPN structure is implicit in SL datasets that used multiple annotators per instance but avoided the multi-principal structure by voting over final labels (Deng et al., 2009; Krizhevsky, 2009; Everingham et al., 2010), and in LLM alignment techniques that elicit annotators preferences over outputs (Ouyang et al., 2022; Rafailov et al., 2023). In general, *MPN models are closely related to multi-objective optimization problems (Roijers et al., 2013; Rădulescu et al., 2020; Röpke et al., 2023), but the *MPN structure highlights the common agency problem.