Decision Analysis, Muddling-Through, and Machine Learning for Managing Large-Scale Uncertain Risks

Louis Anthony Cox, Jr.²

The George Washington University Regulatory Studies Center

ABSTRACT

When is a comprehensive decision analysis the most effective way to choose what to do next, and when might less knowledge-intensive processes – especially, learning as one goes – be more useful? Managing large-scale, geographically distributed, and long-term risks arising from diverse underlying causes ranging from poverty to underinvestment in protecting against natural hazards to failures of critical infrastructure networks and vulnerable sociotechnical, economic, and financial facilities and systems poses formidable challenges for any theory of effective social decision-making. Different affected organizations, populations, communities, individuals, and

---

¹ This working paper reflects the views of the author, and does not represent an official position of the GW Regulatory Studies Center or the George Washington University. The Center’s policy on research integrity is available at [http://regulatorystudies.columbian.gwu.edu/policy-research-integrity](http://regulatorystudies.columbian.gwu.edu/policy-research-integrity).

² Tony Cox is President of Cox Associates, a Denver-based applied research company specializing in quantitative risk analysis, causal modeling, advanced analytics, and operations research. He is a Visiting Scholar at The George Washington University Regulatory Studies Center. Tony is an Edelman Laureate of the Institute for Operations Research and Management Science (INFORMS); a member of the National Academy of Engineering (NAE); a past member of the National Academies’ Board on Mathematical Sciences and their Applications, and a Fellow of the Society for Risk Analysis (SRA).
thought leaders can perceive risks, opportunities, and desirable responses to them very differently. Participants may have different and rapidly evolving local information, goals and priorities; perceive different opportunities and urgencies for actions at any time; and be differently aware of how their actions affect each other through side effects and externalities. Six decades ago, political economist Charles Lindblom viewed theories of “rational-comprehensive decision-making” such as decision analysis and statistical decision theory as utterly impracticable for guiding policies to manage such realistically complex situations. He proposed incremental learning and improvement, or “muddling through,” instead as both a positive and a normative theory of bureaucratic decision making. But sparse, delayed, uncertain and incomplete feedback undermines the effectiveness of collective learning while muddling through, even if all participant incentives are aligned; it is no panacea. We consider how recent insights from machine learning – especially, deep multi-agent reinforcement learning – can be used to formalize several aspects of muddling through. These insights suggest principles for improving human organizational decision-making. Deep learning principles adapted for human use can not only help participants in different levels of government or control hierarchies to manage some large-scale distributed risks better, but they also show how rational-comprehensive decision analysis and incremental learning and improvement can be reconciled and synthesized, making it unnecessary to choose between them.

1. **Introduction**

The normative theory of rational choice is well developed³ and has often been proposed as a prescriptive model for how societal decision-makers – including regulators, planners, bureaucrats, and policy makers – should in principle make decisions in the public interest.⁴ Prescriptive decision analysis typically involves multiple steps such as the following⁵:

1. **Identify alternative feasible choices**, decision rules, or courses of actions. This “choice set,” or set of decision alternatives, is often specified implicitly via constraints on the allowed values of decision variables, such as quantities of limited resources available to be allocated, or precedence constraints specifying which activities must be completed before others can be undertaken.
2. **Identify preferences and value trade-offs** for possible outcomes. These may be formally represented via a (possibly multi-attribute) von Neumann-Morgenstern utility function.⁶
3. **Model the probabilities of different outcomes for each choice** (e.g., its risk profile); and

³ Ramsey 1926; von Neumann and Morgenstern 1944; Savage 1954; Luce and Raiffa 1957; Raiffa 1968.
⁴ e.g., Keeney and Raiffa 1974; Edwards et al. 2007.
⁵ Raiffa 1968; Clemen and Reilly 2014; Howard and Abbas 2015.
⁶ Keeney and Raiffa 1974.
4. **Optimize choices** subject to feasibility constraints (e.g., on available time, budget, or limited resources) to identify and recommend a feasible choice that maximizes expected utility of outcomes.

These steps are all well-established parts of prescriptive decision analysis for a single decision-maker. Extension of this basic theory allows for calculation of value of information (VOI) and optimization of when to take actions vs. collect more information, as well as various forms of robust optimization when outcome probabilities for some choices are uncertain or unknown. The field of decision analysis has flourished in recent decades by extending and applying these principles to a variety of interesting and worthwhile decision problems.

In 1957, political economist Charles Lindblom of Yale University pointed out that almost none of these aspects of normative decision theory can be applied in practice to the decisions and uncertainties faced by real government decision-makers, or by decision-makers in other bureaucracies. First, preferences and value trade-offs may be unknown and difficult or impossible to articulate, quantify, and justify even by those who might wish to do so. If a bureaucrat were to specify them – perhaps using multi-attribute utility theory (MAUT) or multi-criteria decision-making (MCDM) decision aids and probability and utility elicitation methods developed in decision analysis in the decades following Lindblom’s essay – others might disagree with the results. Indeed, even the person who developed them might disagree with them if asked to do the same exercise again later, or if asked using different words or elicitation protocols. Lindblom wrote, “Typically the administrator chooses – and must choose – directly among policies in which [different] values are combined in different ways. He cannot first clarify his values and then choose among policies,” as MAUT prescribes. Preferences and value trade-offs aside, simply identifying the possible outcomes for each feasible choice may be a non-starter in applications where the number of possible choices is immense or possible outcomes are unknown. Even the set of feasible alternative choices may not be known *a priori*, but may require creativity and effort to develop. Outcome probabilities for different choices are often unknown (*a fortiori* if the possible outcomes are themselves unknown). Although robust optimization, uncertainty sets, coherent risk measures, and other tools of modern risk analysis and operations research help to address some of these difficulties, the impracticality of a full decision-analytic approach to many real-world problems still seems clear to many decision makers (although perhaps not always to decision analysts).

In addition, real-world bureaucratic and organizational decisions are almost never made by a single decision-maker. Even an ideal decision analysis using a coherent set of subjective probabilities and utilities would not necessarily represent the beliefs and risk attitudes of many of those affected.

---

7 Howard and Abbas 2016; Raiffa 1968.
8 Edwards et al. 2007; Clemen and Reilly 2014; Howard and Abbas 2015.
9 Clemen and Reilly 2014.
by or accountable for the decision. An impressive array of impossibility results in collective choice theory and game theory indicates the difficulty of passing from single-person to multi-person normative decision theory in all but the most special cases. Table 1 summarizes some of these results, although it is not self-contained and seeks only to give an idea of key results. (For example, the quoted Myerson-Satterthwaite impossibility theorem uses concepts of Bayesian incentive-compatibility (reporting one’s true valuation is a Bayesian Nash equilibrium, i.e., neither player expects to gain by unilaterally misrepresenting his preferences), individual rationality (essentially, that a rational player would voluntarily participate), and \textit{ex post} efficiency (the player who values the object more ends up with it) that are not explained in the table.) The cited references provide precise statements, detailed explanations, and insightful discussions of limitations and extensions of these impossibility results. But the main point is that no collective choice procedure satisfies all of the desired properties that have been proposed, such as Pareto efficiency or unanimity (the collective choice procedure should not select one alternative if everyone prefers a different one), separate aggregation of beliefs and of preferences, various concepts of freedom (e.g., that participation should be voluntary and society should accept at least some individual preferences) and so forth. Moreover, individuals with private information, even if it is only about their own preferences and beliefs, often have incentives to strategically misrepresent itin collective choice procedures, undermining ability to identify Pareto-efficient collective outcomes. In the face of such challenges and inconsistencies of desiderata, normative decision analysis cannot easily be extended to collective choices made by multiple decision-makers.

Rather than seeking to extend or refine normative decision analysis to overcome what he perceived as its fatal practical limitations for large-scale, multi-person organizational decision-making over time, Lindblom instead described a method of \textit{successive limited comparisons} that he contrasts with the “rational-comprehensive” normative approach favored in decision analysis, operations research, and optimal control engineering. The rational-comprehensive approach to individual, organizational, or societal decision-making seeks to solve decision optimization problems such as

$$\max_{a \in A} R(a)$$

(1)

where

- $a$ is a decision variable or policy (perhaps a vector or a time series of inputs provided to a controlled system, or a feedback control rule for mapping observations to actions)
- $A$ is the set of feasible alternative decisions (the “choice set”)
- $R(a)$ is the reward or expected utility from choosing $a$. In decision analysis, the reward function $R$ to be maximized (also called an objective function, expected utility function, or payoff function) is usually assumed to be known. In statistical design of experiments and response surface analysis and in machine learning, it may have to be discovered. If the reward received depends both on the decision-maker’s choice $a$ and also on other variables not controlled by the decision-maker, collectively referred to as the \textit{state} and
modeled as a random variable in decision analysis, then $R(a)$ is the expected reward or utility from choosing $a$ given the probability distribution of $s$. When there are many players, $R$ is often taken to be a weighted sum of individual utility functions. This can be motivated by considerations of Pareto efficiency and both individual and collective rationality (i.e., assuming that individual and collective preferences should both satisfy the von Neumann-Morgenstern axioms), and many impossibility results can be avoided by dropping the unanimity requirement that collective preferences among alternatives must agree with shared individual preferences, if they are based on conflicting beliefs:

- $\max_{a \in A}$ indicates that an act $a$ in $A$ is to be selected to maximize $R(a)$.

### Table 1: Impossibility Results for Rational Collective Decision-Making

<table>
<thead>
<tr>
<th>Impossibility Result</th>
<th>Main Idea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrow’s Theorem$^{12}$</td>
<td>No procedure for aggregating individual preference orderings for three or more alternative collective choices to obtain a coherent (transitive) collective preference ordering satisfies the following four properties:</td>
</tr>
<tr>
<td></td>
<td>- Pareto-efficiency (not everyone prefers a different choice)</td>
</tr>
<tr>
<td></td>
<td>- Non-dictatorship: The collective ordering is responsive to the preference orderings of more than one participant</td>
</tr>
<tr>
<td></td>
<td>- Unrestricted preferences: Individual preference orderings are unrestricted. (This is usually referred to as “unrestricted domain.”)</td>
</tr>
<tr>
<td></td>
<td>- The collective ranking of any pair of alternatives depends only on their rankings by individuals (it is “independent of irrelevant alternatives”)</td>
</tr>
<tr>
<td>Sen’s Liberal Paradox$^{13}$</td>
<td>No transitive aggregation rule satisfies Pareto efficiency, unrestricted preferences and minimal liberalism, a special case of non-dictatorship stating that each of at least two individuals can choose between at least two alternatives without being overruled by the collective preference ordering.</td>
</tr>
<tr>
<td>Gibbard-Satterthwaite Theorem$^{14}$</td>
<td>Any non-dictatorial choice function (rule or procedure mapping individual statements of preferences to a collective decision) used by players with unrestricted preferences to choose among</td>
</tr>
</tbody>
</table>

---

$^{11}$ Gilboa et al. 2004.
$^{12}$ Arrow 1963.
$^{13}$ Sen 1970.
| Impossibility of Bayesian group decision making with separate aggregation of beliefs and values\(^{15}\) | “It is assumed that the group's belief depends only on individual beliefs and the group's values only on individual values, that the belief aggregation procedure respects unanimity, and that the entire procedure guarantees Pareto optimality. We prove that only trivial (dictatorial) aggregation procedures for beliefs are possible.” |
| Myerson-Satterthwaite impossibility theorem\(^{16}\) | “We consider bargaining problems between one buyer and one seller for a single object. The seller’s valuation and the buyer’s valuation for the object are assumed to be independent random variables, and each individual’s valuation is unknown to the other. We characterize the set of allocation mechanisms that are Bayesian incentive compatible and individually rational, and show the general impossibility of *ex post* efficient mechanisms without outside subsidies.” |
| Inconsistency of intergenerational equity and Pareto efficiency\(^{17}\) | “It has been known that, in aggregating infinite utility streams, there does not exist any social welfare function, which satisfies the axioms of Pareto, intergenerational equity, and continuity. We show that the impossibility result persists even without imposing the continuity axiom…” |
| Impossibility of a Paretian Rational\(^{18}\) | “If a group takes a collective decision on the basis of separately aggregated group judgments on the probabilities of independent events, there may not exist any anonymous aggregation rule that respects individuals’ unanimous outcome preferences at all profiles of beliefs.” |

Lindblom wrote that “the attention given to, and successes enjoyed by operations research, statistical decision theory, and systems analysis” have strengthened a “tendency to describe policy formulation even for complex problems as though it followed [this] approach,” emphasizing “clarity of objective, explicitness of evaluation, a high degree of comprehensiveness of overview, and, wherever possible, quantification of values for mathematical analysis. But these advanced procedures remain largely the appropriate techniques of relatively small-scale problem-solving where the total number of variables to be considered is small and value problems restricted.”

\(^{15}\) Zeckhauser and Hylland 1979.  
\(^{16}\) Myerson and Satterthwaite 1983.  
\(^{17}\) Basu and Mitra 2003.  
\(^{18}\) Nehring 2007.
In contrast, for large-scale real-world decision problems faced by most bureaucracies, Lindblom considers the rational-comprehensive approach in equation (1) to be impracticable because reward function $R$ is not known or agreed to; choice set $A$ is too large to enumerate or search effectively, or possibly is unknown and costly to develop; and no single centralized authority is capable of, authorized to, or accountable for identifying and implementing the best choice in $A$. In practice, instead of clarifying values and objectives in advance, goals and actions to achieve them are selected together as opportunities arise. The test of a “good” policy is not that it is the best means to desired ends, or that it maximizes some measure of expected utility or collective welfare, but that people will agree to it (possibly for different, and perhaps conflicting, private reasons). Important possible outcomes, feasible alternative policies, and affected values and trade-offs are neglected in favor of relatively simple comparisons between the current policy and a proposed incremental modification of it. A succession of such modifications may, if all goes well, produce gradually improving policies; this is the process that Lindblom refers to as successive limited comparisons, or, more colloquially, as muddling through. He states that “Making policy is at best a very rough process. Neither social scientists, nor politicians, nor public administrators yet know enough about the social world to avoid repeated error in predicting the consequences of policy moves. A wise policy maker consequently expects that his policies will achieve only part of what he hopes and at the same time will produce unanticipated consequences that he would have preferred to avoid. If he proceeds through a succession of incremental changes, he avoids serious lasting mistakes in several ways” including learning from experience and being able to correct missteps fairly quickly. Of course, this view is optimistic if a single misstep could lead to disaster, ruin, or the destruction of the decision-making organizations, but Lindblom does not dwell on these grim possibilities. In modeling and evaluating the muddling through approach more formally, however, we will have to consider possibilities for safe learning, i.e., surviving and avoiding disastrous decisions during learning.\footnote{Garcia and Fernandez 2015.}

Lindblom proposes that muddling through has advantages not only as a descriptive theory of bureaucratic decision-making, but also as a normative one: “Why then bother to describe the method in all of the above detail? Because it is in fact a common method of policy formulation and is, for complex problems, the principal reliance of administrators as well as of other policy analysts. And because it will be superior to any other decision-making method available for complex problems in many circumstances, certainly superior to a futile attempt at superhuman comprehensiveness.” In short, muddling through by successive incremental adjustments of policy is proposed as both more desirable and more widely practiced than the rational-comprehensive approach.

Six decades have passed since Lindblom’s essay. Revolutions have occurred in computer science, game theory, collective choice theory, automated and adaptive control, artificial intelligence,
robust optimization and risk analysis, machine learning, computational statistics and data science, and the intersection of these fields with political economy, law-and-economics, and management science. Operations research has matured and improved, in part by developing algorithms to solve large-scale versions of optimization problem (1). In artificial intelligence and machine learning, algorithms are increasingly being developed and tested to manage swarms of drones, autonomous vehicles, traffic routing in telecommunications networks or urban traffic grids, loads and capacity in power networks, robot soccer teams, military logistics and engagements, and other settings where autonomy and control are distributed among multiple agents at different locations with different information, resources, and opportunities for action. It is timely to reexamine the extent to which Lindblom’s critique of rational-comprehensive techniques for risk management decision support still applies; the extent to which the ferment of ideas and technical developments in artificial intelligence and other fields dealing with multi-agent control has overcome his objections; how both the strengths and the limitations of muddling through can be understood better, and the technique applied more successfully, in light of progress since 1957; and whether there are circumstances in which muddling through provides a viable alternative or complement to decision analysis. The following sections undertake such a reexamination.

Practical motivations abound. Examples of large-scale risk management challenges in which effective action requires coordination among many agents, despite highly uncertain futures and consequences of current choices, include eradicating or containing the spread of geographically distributed diseases (e.g., polio, mad cow disease, rabies); controlling the emergence and dissemination of antibiotic-resistant bacteria within and between countries; reducing air pollution health effects; and taking actions to reduce risks from climate change. We will return to the antibiotic resistance and air pollution examples later to illustrate some practical points about the challenges of learning to muddle through successfully. Each of these and countless other examples requires effective coordination among multiple actors (“agents” in artificial intelligence parlance) at various levels of different bureaucracies, from hospitals to local, state, and national governments. Each requires learning as well as doing, since not enough is currently known to predict the long-run effects of proposed policy interventions and innovations with much accuracy and confidence. We consider how an improved science of muddling through might help to develop practical solutions for better managing such large-scale, geographically distributed, and long-term risks.

II. Developments in Rational-Comprehensive Models of Decision-Making

The decision-analytic foundations of the rational-comprehensive model of individual decision-making are the same today as when Lindblom wrote, with expected utility maximization remaining as the dominant normative paradigm, but many new insights, principles, and techniques have been developed for situations where the elements of a decision problem – especially the conditional
probabilities of different outcomes and utilities when different choices are made – are initially unknown. In addition, there have been dramatic advances in theories and methods for multi-agent decision-making. This section surveys some important developments in normative models of rational-comprehensive decision-making since Lindblom’s essay.

An individual, team, organization, or artificial intelligence that repeatedly makes decisions to achieve some overall purposes or goals must repeatedly decide what to do next – e.g., what subgoals or tasks to undertake next – and how to do it. In a bureaucracy with at least some top-down or centralized coordination and control, this may require developing and sharing a plan and assigning roles, tasks, or subgoals to the different agents working to execute it. In an uncertain and complex world, execution of the different parts of a plan may or may not succeed. Agents must then evaluate and communicate their progress or lack of it, and perhaps their expectations and pertinent local information, so that the current plan or its pieces can be modified or replaced if it turns out that conditions warrant doing so. On the other hand, agents may also discover unexpected local opportunities or constraints that justify updating their own immediate goals and commitments to each other; again, information discovered by one agent may be useful to others, or to a centralized planner, if it can be used to improve plans or their execution. In teams with no central coordinator, such as robot soccer teams of cooperating autonomous agents, cooperating swarms of drones, or search-and-rescue teams with autonomous agents and limited communication, the agents may have to infer and adapt to each other’s plans on the fly as they observe each other’s behaviors and messages. However, our focus remains on bureaucracies or other organizations where policies are formulated and adapted via muddling through, by successive rounds of modifying previous policies. Who is able to propose what when, exactly how decisions are made about which proposals to adopt (and how and whether these decisions, once made, are implemented, monitored, and enforced), and how these changes and their consequences are linked to incentives and rewards for those participating in policy making and administration, all vary widely across organizations.

In the face of such complexities, the simple prescriptive model of optimization-based rational-comprehensive decision-making in (1) and the techniques for solving them that were available when Lindblom wrote been generalized and extended in the following ways.

Multi-Agent Decision Theories and Models

- **Non-cooperative game theory** replaces the reward function $R(a)$ in (1) with a set of reward functions (also called “payoff functions”), one for each participant (called a “player” or “agent”). Each player has its own choice set of feasible alternatives to choose among, often called strategies in game theory, or policies in decision analysis, machine

---

21 Luce and Raiffa1957.
learning, and artificial intelligence. In very simple games, these alternatives may be actions or behaviors, such as which move to make next in a board game, or which behavior (e.g., submit vs. fight) to adopt in a confrontation. More generally, they are decision rules (also called policies) determining what to do next based on what an agent has observed so far and on the choices that are available. In non-cooperative game theory, each player seeks to choose a strategy from its choice set to maximize its own reward (i.e., expected utility or payoff), given the choices of the other players. Some key concepts and insights from game theory are as follows.

- Optimization problem (1) for a single decision-maker is generalized to a set of interdependent optimization problems, one for each player. Player \(i\) now seeks to choose \(a_i\) from \(A_i\) to maximize \(R_i(a_i, a_i')\), where \(a_i\) denotes the strategy selected from \(A_i\) by player \(i\); \(a_i'\) denotes all the strategies selected by the other players; and \(R_i(a_i, a_i')\) is the reward to player \(i\) from choosing strategy \(a_i\) when the other players choose \(a_i'\). There is no single social welfare or public interest to be maximized. Rather, each player seeks to act to maximize its own reward, given the actions of the rest.

- A Nash equilibrium is a set of choices such that no player can improve its own reward by unilaterally modifying its own choice, given the choices of the other players. Each player’s choice is a best response to the choices of the rest.

- A set of choices by the players is Pareto-efficient if no other set of choices would give all players equal or greater rewards, and at least some of them greater rewards.

- A fundamental challenge for cooperation and coordination among agents, illustrated by the notorious Prisoner’s Dilemma and Tragedy of the Commons games, is that sometimes no Nash equilibrium choice is Pareto-efficient. In the motivating examples of managing air pollution, antibiotic resistance, and climate change, a common challenge is that each player benefits if everyone else exercises restraint to avoid making the current problem worse, but each player also maximizes its own benefits by being unrestrained itself, whatever the other players are doing. In such cases, the Nash equilibrium is that no one exercises self-restraint, even though all would gain if all would do so.

- A variety of “folk theorems” of game theory prove that both Pareto efficiency and multi-period versions of Nash equilibrium can be achieved if players are sufficiently patient (i.e., they do not discount delayed rewards too steeply) in repeated games with discounted rewards and uncertain time horizons, where the players have a chance to observe each other’s behaviors and make choices repeatedly over time. The trick is to have players make choices that punish those who do not cooperate in sustaining a Pareto-efficient outcome.\(^{22}\)

\(^{22}\) Fudenberg and Maskin 1986; Fudenberg et al. 1994; Hörner, J. and Olszewski 2006.
Cooperative game theory further generalizes the multi-player choice problem by allowing players to form coalitions and to bargain or negotiate with each other. For example, in the treaty participation game model of international cooperation (or lack of it) to limit emissions in hopes of limiting undesired climate change, a coalition of signatories might choose emissions levels to maximize their collective benefits, while non-signatories choose emissions levels to maximize their individual benefits. The final levels of cooperation and emissions achieved in multistage games of coalition formation and decision-making about emissions depend on factors such as whether coalitions, once formed, are exclusive; whether players (e.g., countries) can make and enforce conditional agreements such as that some will reduce their emissions more if and only if others do; whether binding commitments can be made and enforced; how steeply participants discount future rewards and penalties compared to current ones; and whether the timing of catastrophic consequences from failure to muster sufficient cooperation is known or uncertain.

Team theory focuses on design of communication and agent decision rules (and, in some versions, on allocation of limited resources among the agents) for the special case of cooperating agents in an organization where all of the agents have identical preferences and goals. That is, they all seek to maximize the same reward function of their joint choices, but local observations, actions, and communications are costly. Team theory has been applied to distributed control of systems by agents with sensors and actuators at different locations, as well as to organizational design, design of compensation systems, and dynamic allocation of tasks, roles, and responsibilities within teams of cooperating agents.

Mechanism design: Institutions, social and moral norms, legal constraints and liabilities, regulations and their enforcement, wages and contractual incentives, outcome-sharing rules in principal-agent relationships and investment syndicates, and reputations in repeated transactions and long-term relationships all help to shape the rewards (positive or negative) and feedback that players receive for their choices and behaviors. Game theory studies how agents make choices in response to incentives. Mechanism design theory studies the inverse problem of how to design incentives, or the rules determining rewards in the games in which agents participate, to elicit choices that satisfy desired properties. These may include Pareto efficiency, self-enforcing stability (e.g., Nash equilibrium and its multi-period extensions), implementability using information that can actually be obtained and incentives (e.g., payments) that can actually be provided, and voluntary participation. The impossibility theorems in Table 1 illustrate the challenges of

---

25 Marschak and Radner 1972.
successful mechanism design, but many positive results are available when preferences satisfy some restrictions (e.g., risk neutrality and “quasilinear preferences” with utility linear in money).

• **Organizational design and law-and-economics:** Within bureaucracies and other hierarchical organizations (e.g., principal-agent relationships), as well as in the more specialized contexts of designing contracts and auctions, the theory of mechanism design can be applied to design incentive systems to promote revelation of local information, elicit desired behaviors despite private information, and optimize delegation and tradeoffs between centralization and decentralization, taking into account costs of communication, monitoring, and control and inefficiencies due to remaining private information. As a prominent application of the mechanism design perspective, the modern theory of law and economics explains how systems of laws establishing tort liability rules for hazardous activities, remedies for breach of contracts, property rights to internalize externalities, product liability and implicit warranty principles, and so forth can be designed to maximize the expected net economic benefit from voluntary transactions, usually assuming risk-neutral participants with quasilinear preferences. Practical designs that explain many aspects of observed legal practice account for market imperfections such as private and asymmetric information (e.g., a consumer may not know how much care a manufacturer has taken to keep a product safe, or the manufacturer may not know how much care the consumer will exercise in using the product safely), costs of litigation, misperceptions of risk by buyers, and incentives for socially valuable research and disclosure of information by sellers.

**Modern Algorithms for Single- and Multi-Agent Decision-Making**

The intersection of computer science with decision models and algorithms has tremendously advanced the design and practical application of algorithms for solving large-scale single-person and team decision optimization problems, as well as games and collective choice problems, in recent decades. Current state-of-the-art algorithms are briefly described next.

• **Monte Carlo Tree Search (MCTS).** Decision trees showing possible sequences of actions (choice nodes) and uncertainty resolutions (chance nodes, with probabilities for each branch) leading to rewards (utilities) at the ends (leaf nodes) of the tree are perhaps the best known rational-comprehensive models of normative decision analysis for small problems. In practice, combinatorial explosion renders them “bushy messes” for larger problems, motivating the use of influence diagrams and other more efficient

---

27 Mookherjee 2006.
29 Raiffa 1968.
representations of decision problems\textsuperscript{30}. However, both decision trees and their
generalizations to multiple players, i.e., game trees\textsuperscript{31} have recently been made
dramatically more useful for large problems by \textit{Monte Carlo Tree Search} (MCTS)
algorithms\textsuperscript{32}. These apply Monte Carlo sampling and evaluation of possible future paths
and rewards to avoid enumerating all possibilities. MCTS algorithms thus decouple
“rational” decision-making, based on optimizing current decisions based on predicted
future reward probabilities, from “comprehensive” modeling of the causal relationship
between choices and reward probabilities by selecting only the most promising choice
nodes in a tree for further simulation and evaluation. MCTS can be combined with
reinforcement learning (RL) techniques discussed next\textsuperscript{33} and applied to more general
settings, such as those in which it is costly to observe the reward\textsuperscript{34}, as is the case for
many social policy interventions.

- \textit{Reinforcement learning (RL) of high-reward policies through trial and error learning}\textsuperscript{35}.
  Decision-makers (agents) often initially do not know how their choices affect reward
  probabilities, but must discover what works from experience. Let the true expected value
  starting in state $s$ and acting optimally thereafter be denoted by an (initially unknown)
  \textit{value function}, $V(s)$. Let $Q(a, s)$ denote an estimate of the value from taking each feasible
  action $a$ when in each state $s$ and then acting optimally (e.g., to maximize the discounted
  sum of future rewards) ever after. The initial estimates of these values may be random
  guesses, but they are updated in light of experience by adjusting current estimates by an
  amount proportional to the difference between expected and experienced rewards. The
  constant of proportionality is interpreted as the \textit{learning rate}. For example, $Q$-\textit{learning}
  uses the current estimate $Q(a, s)$ to select which action to take next in the current state $s$.
  Then the resulting reward is used to update the estimate of $Q(a, s)$ based on the difference
  between estimated and observed rewards. In many settings, estimated $Q(a, s)$ values
  converge and the policy of selecting $a$ to maximize $Q(a, s)$ is then the optimal policy,
  while the estimated value of $Q(a, s)$ when that policy is used is the true value function,
  $V(s)$. This procedure is similar to value iteration in classical stochastic dynamic
  programming, but without the requirement that the reward function and state transition
  probabilities be initially known. It converges to yield optimal policies under certain
  conditions for Markov decision processes (MDPs), in which the actions taken affect
  next-state probabilities as well as probability distributions of current rewards) (Krishnamurthy
  2015). The main conditions are that learning rates be kept small enough and that the
  MDPs are ergodic, involving no irreversible choices or fatal outcomes that would limit or

\textsuperscript{30} Howard 1988.
\textsuperscript{31} Luce and Raiffa 1957.
\textsuperscript{32} Munos 2014; Silver et al. 2016, 2018.
\textsuperscript{33} Vodopivec et al. 2017.
\textsuperscript{34} Schulze and Evans 2018.
prevent future exploration and adaptation (Bloembergen et al. 2015; Krishnamurthy 2015; Xu et al. 2017).

- **RL using policy gradient algorithms.** RL can also be based on algorithms that emphasize adjusting policies directly rather than estimating values for different actions. As usual, a policy in RL is a decision rule mapping observations (e.g., the current state) to actions. In most RL algorithms, however, this mapping is randomized: thus, a policy RL specifies the probability of taking each feasible action when in each state (or, more generally, given current information, which may include imperfect observations of the current state). Policies are updated to favor selecting actions with higher expected values. The tension between exploring further in hopes of finding a more valuable policy and exploiting what has been learned so far by selecting the actions with the highest expected values is managed carefully by choosing action-selection probabilities to avoid premature convergence to sub-optimal policies. For example, a simple and effective policy in many settings is to select each action with a probability equal to the currently estimated probability that it is the best (value-maximizing) action; this is called Thompson sampling. Such randomized sampling schemes prevent jumping to possibly erroneous conclusions about what works best in clinical trials and similar sequential decision optimization settings. Adjustments of policies continue until expected and experienced average rewards no longer differ. For large classes of adaptive decision problems under uncertainty, the policies arrived at by such successive incremental adjustments are the optimal policies that would be obtained by classical operations research methods, such as solving the Bellman stochastic dynamic programming recursion. For example, RL algorithms eventually converge to optimal policies for ergodic MDPs with fixed but initially unknown state transition and reward conditional probabilities if step sizes are kept small enough. A major advantage of RL algorithms over classical methods, however, is that they discover optimal policies without the perhaps unrealistic a priori knowledge requirements of classical methods, such as knowledge of how actions affect state transition probabilities and conditional probability distributions of rewards.

- **Enhanced RL:** RL has been implemented via several different algorithms for iterative selection of actions, observation of resulting rewards, and improvement of value estimates and policies. Table 2 lists important refinements and enhancements used in practice to make learning quicker and more robust to data limitations. We will return to several of these ideas later.

- **Safe RL:** For real-world applications, it is often vital to assure that learning processes are safe, meaning that the learning algorithms will not cause accidents, disasters, or large

---

36 Schulze and Evans 2018.
37 Villar et al. 2915.
38 Bloembergen et al. 2015; Krishnamurthy 2015; Xu et al. 2017.
39 Sutton and Barto 2018.
losses while learning. Table 3 summarizes several methods for safe learning. These have proved effective in applications ranging from learning to control helicopters and quadcopters (e.g., allowing them to hover or navigate safely in cluttered environments) to learning to manage power grids and other networked infrastructures, without risking costly accidents and failures during learning. Safe learning can help to assure that an optimal policy is eventually reached even for non-ergodic MDPs by avoiding irreversibly catastrophic decisions.

- **Multi-agent reinforcement learning (MARL).** Multiple agents acting, learning, and perhaps communicating in parallel how to control a system or accomplish a task can greatly increase speed of learning and average rewards generated per unit time, under certain conditions. Table 4 summarizes several different variations and extensions of MARL, including alternative communication and control architectures and results.

Table 2: Some enhancements to Reinforcement Learning (RL) algorithms

<table>
<thead>
<tr>
<th>Enhancement</th>
<th>Main Ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy gradient RL algorithms</td>
<td>Directly modify policies, without first estimating a value function for the states, by estimating the gradient (slope) of the reward as a function of policy parameters and adjust those parameters incrementally to ascend the estimated slope.</td>
</tr>
<tr>
<td>Actor-critic architectures</td>
<td>Interpret the policy at any time as an “actor” and the value function as a “critic” that evaluates how well the current policy is working. Separating these two roles helps to speed convergence.</td>
</tr>
<tr>
<td>Model-based RL</td>
<td>Fit statistical models of reward probabilities and state transition probabilities to observed state-act-reward-next-state data. Use the models to speed learning of high-reward policies (if the models are usefully accurate).</td>
</tr>
<tr>
<td>Model-free RL</td>
<td>Use empirically observed rewards to estimate state or action value functions (via iteratively updated ( Q ) values). Powerful statistical and machine learning techniques for approximating unknown functions from data, such as deep neural networks, can obtain most of the advantages of model-based RL while avoiding the potential pitfalls from using incorrect models.</td>
</tr>
</tbody>
</table>

41 Arulkumran et al. 2017.
42 Grondman et al. 2012.
43 Clavera et al. 2018.
44 Mnih et al. 2015; Andrychowicz et al. 2018.
<table>
<thead>
<tr>
<th>Technique</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reward shaping</td>
<td>Modify the original reward function received from the environment to encourage quicker learning and discovery of better policies.</td>
</tr>
<tr>
<td>Experience replay</td>
<td>Use Monte Carlo simulation from frequency distributions of past experiences (e.g., state-action-reward-next state sequences) to reduce computational burden and augment sparse training data.</td>
</tr>
<tr>
<td>Deep learning control of the learning rate</td>
<td>Use deep learning neural networks to automatically adjust the learning rate parameter using an actor-critic architecture in which one neural network adjusts the parameter and another provides feedback on how well the adjustments appear to be working.</td>
</tr>
<tr>
<td>Meta-learning</td>
<td>Estimate crude high-level models of rewards and value functions relatively rapidly. Refine and improve them and use them to guide actions via RL as new observations are made. Such a hierarchy of modeling allows relatively rapid and effective adaptation to new conditions in non-stationary, including graceful compensation for and recovery from partial system failures.</td>
</tr>
<tr>
<td>Inverse RL and imitation learning.</td>
<td>Use observed data on state and action sequences leading to success or failure in a task to infer successful policies for choosing actions to take in each state to accomplish it successfully. This makes it possible for agents to learn quickly from humans or other more experienced and higher-performing agents how to do complex tasks.</td>
</tr>
<tr>
<td>Hybrids of above techniques</td>
<td>Example: Interleaving updates of the estimated value function with sampling from the experience replay buffer and adjustment of policies to increase expected reward (“policy gradient ascent” for rewards or “policy gradient descent” for losses, using a step size determined by the current learning rate parameter).</td>
</tr>
</tbody>
</table>

46 Andrychowicz et al. 2018.
47 Xu et al. 2017.
48 Lemke et al. 2015; Clavira et al. 2018.
49 Shiarlis et al. 2016.
Table 3: Some principles for safe learning, i.e., learning without risking catastrophic failures

<table>
<thead>
<tr>
<th>Safe learning principles</th>
<th>Main Ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk-sensitive learning and control</td>
<td>Modify the reward function to consider variance in return; probabilities of ruin or large loss, such as crash of an autonomous vehicle; and risk-sensitive control policies.</td>
</tr>
<tr>
<td>Imitation learning with safe instruction</td>
<td>Use imitation learning from demonstrations supplied by instructors to assure that only safe examples are imitated.</td>
</tr>
<tr>
<td>Knowledge-based constraints on exploration</td>
<td>Use knowledge-based constraints supplied by instructors to assure that only safe changes in policies are explored during learning.</td>
</tr>
<tr>
<td>Maintain system stability while learning and exploring modified policies</td>
<td>Apply feedback control theory for dynamic systems to maintain stability of the system while collecting data. Use the collected data to learn to improve control performance and to expand the safe region of the state space, i.e., the set of states for which safe control policies are available. Keeping changes in control policies small enough to avoid destabilizing the system while learning is effective for systems that are known to have well-behaved dynamics, without large (e.g., discontinuous jump) responses to small changes in controls.</td>
</tr>
<tr>
<td>Use model uncertainty to constrain exploration</td>
<td>Create uncertainty zones around regions of potentially high loss (e.g., around pedestrians with unpredictable behaviors) based on model uncertainty estimates, and avoid them during learning.</td>
</tr>
<tr>
<td>Safe policy improvement using a known safe policy as default when model uncertainty is high</td>
<td>Engage in safe policy improvement by using known safe (i.e., catastrophe-avoiding) default policies when model uncertainty about effects of changing the policy is high. Explore for possible improvements in policies when model uncertainty is low.</td>
</tr>
<tr>
<td>Safe policy improvement using statistical confidence bounds to limit the risk from policy modifications</td>
<td>Use statistical confidence bounds (e.g., derived from importance sampling and probability inequalities) for performance of modified policies to avoid those that pose unacceptable risks.</td>
</tr>
</tbody>
</table>

50 Garcia and Fernandez 2015.  
51 Ibid.  
52 Ibid.  
54 Lütjens et al. 2018.  
56 Thomas et al. 2015.
### Table 4: Some MARL variations and extensions

<table>
<thead>
<tr>
<th>Setting</th>
<th>Main ideas, results, and applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARL for non-cooperative stochastic games</td>
<td>Convergence to Nash equilibria occurs under certain conditions if each agent uses RL and manages its learning rate appropriately. (However, Nash equilibria may be Pareto-efficient.)</td>
</tr>
<tr>
<td>Collective choice MARL</td>
<td>Agents initially know only their own preferences. They negotiate by proposing joint actions to each other to improve their own payoffs. Accepted proposals are binding and generate mutual gains. This cooperative negotiation leads to Pareto-superior outcomes than non-cooperative MARL in many games.</td>
</tr>
<tr>
<td>MARL for teams without communication among agents</td>
<td>Teams of cooperating agents with the same goal (i.e., cooperating to maximize the same reward function) can learn to behave effectively in many settings even without explicit communication, by observing, modeling, and adjusting to each other’s behaviors.</td>
</tr>
<tr>
<td>Decentralized MARL for distributed control of a system by a team of cooperating and communicating agents</td>
<td>Decentralized cooperative learning by a team of agents based on explicit communication (e.g., over an unreliable communication network), with agents sharing experiences (data, estimated value functions, or policies), improves learning of distributed control policies to maximize average reward. Applications include control of power grids, mobile sensor networks, and autonomous vehicles.</td>
</tr>
<tr>
<td>Hierarchical MARL (HMARL)</td>
<td>MARL systems with hierarchical organizations of agents, as well as other techniques such as reward shaping, speed convergence to high-reward policies in many settings.</td>
</tr>
<tr>
<td>Decentralized multi-level HMARL</td>
<td>In a multi-level hierarchy of agents, supervisory agents abstract and aggregate information from their subordinates, share it with their peers, pass summaries upward to their own supervisors, and pass supervisory suggestions and constraints on next actions down to their subordinates. This approach has been</td>
</tr>
</tbody>
</table>

---

58 Hu and Gao 2015.
59 Gupta et al., 2017.
60 Zhang et al. 2018.
61 Mannion, 2018.
found to improve convergence of MARL learning in tasks requiring distributed control, such as network routing.\(^{62}\)

<table>
<thead>
<tr>
<th>Two-level HMARL</th>
<th>A central controller coordinates learning among the agents. Local agents manage different parts of a system, such as a supply chain network. They send to the central controller information about their current policies (e.g., represented as deep neural networks for mapping observations to actions) and observations on local costs (e.g., arising from inventory ordering, holding, and stockout costs). The central controller sends feedback to the agents (e.g., weights for the best policies learned so far by each agent) to coordinate their learning. In experimental supply chains, such two-level hierarchical MARL systems discovered policies that substantially reduce costs (e.g., by 80%) compared to the performance of human managers.(^{63})</th>
</tr>
</thead>
</table>

| Hierarchy of tasks assigned to a hierarchy of agents | Hierarchical deep MARL can be used to decompose a learning task into a hierarchy with high-level learning of policies over multi-step goals and low-level controllers learning policies for taking the actions or steps needed to complete those goals. This task decomposition architecture combined with experience replay proved effective for learning high-reward policies in complex and rapidly changing test environments, such as managing a team of cooperating agents in a simulated basketball attack/defense game, even in the presence of sparse and delayed rewards.\(^{64}\) |

MARL algorithms and architectures that incorporate the other advances just discussed – MCTS, RL with various enhancements to speed convergence, safe learning, and communication and control hierarchies such as those in Table 4 – represent the current state-of-the-art in machine learning models and methods for solving large-scale and distributed decision and control problems under uncertainty, including problems with sparse and delayed feedback. Although many MARL algorithms are designed for use with teams of cooperating agents, several also provide performance guarantees for agents engaged in more general game-theoretic interactions. For example, Bowling and Voloso (2001) showed that convergence to Nash equilibria can be achieved in a variety of Markov games (generalizations of MDPs to multiple agents) if each agent uses RL but manages its learning rate to take large steps when the agent’s experienced rewards are less than expected (“learn fast when losing”) and small steps otherwise (when it is “winning” by receiving higher than expected rewards). The resulting WoLF (“win or learn fast”) principle has been incorporated into many subsequent MARL algorithms for cooperative learning. It gives agents

\(^{62}\) Zhang et al. 2008.

\(^{63}\) Fuji et al. 2018.

\(^{64}\) Tang et al. 2018.
who are lagging in learning to contribute to the team’s success time to catch up, while agents who are ahead of the rest continue to explore relatively cautiously (via relatively small incremental adjustment steps) for even better policies.

MARL algorithms have been found to exhibit two valuable performance properties in many decision contexts, although achieving them may require careful tuning of learning rates and exploitation of other ideas in Table 2:

(i) **Convergence**: The learning process converges, in that some or all of agent beliefs (represented by their estimates of value functions and state probabilities), behaviors (represented by agent policies and the choices they generate), or reward distributions eventually stop adjusting. Convergence guarantees typically require individual agents to control their learning rates, e.g., using WoLF or other heuristics, so that they jointly learn to extract reward from the environment, rather than simply adapting to each other.

(ii) **Approximate optimality**: For teams of agents seeking to cooperate to maximize the same reward function in an ergodic Markov game, several MARL algorithms are effective according to various metrics, such as coming close to the optimal cumulative or average reward produced by a single centralized controller solving a very large MDP on behalf of all agents; or achieving small regret, as measured by the difference or ratio of the expected rewards generated by the learning process and the expected rewards that could have been achieved had the optimal policy initially been known and followed.

However, MARL is no panacea for difficult decision problems. These desirable performance properties are guaranteed only for specific (although useful) classes of decision models, such as Markov games (many-player MDPs) under certain assumptions, such as continued survival and unlimited opportunities for exploration and improvement\(^{65}\). How well MARLs work in solving important real-world large-scale decision and control problems under uncertainty, and what combinations of techniques makes them work best, are still being evaluated in a very active applied research literature in applied artificial intelligence, machine learning, and operations research.

In practice, MARL algorithms have been applied successfully to obtain high-reward policies for difficult distributed decision and control problems such as job shop scheduling among multiple agents\(^{66}\); coordination of military force attacks in increasingly large-scale and realistic war game simulations (e.g., StarCraft battles)\(^{67}\); and self-organizing control of swarms of drones to perform missions or to cooperate in choosing locations to obtain full visual coverage of a complex and

---

\(^{65}\) e.g., Zhang et al. 2018.

\(^{66}\) Gabel and Riedmiller 2007.

\(^{67}\) Usunier et al. 2016.
initially unknown environment. Safe MARL (SMARL) and Hierarchical MARL (HMARL) algorithms have demonstrated promising performance in controlling autonomous vehicles and teams of robots performing challenging tasks such as urban search and rescue in complex and uncertain environments, respectively. Such results suggest the potential for MARL principles and their extensions to contribute to improved control of complex distributed systems in important practical business, military, and industrial engineering applications.

III. Implications of Advances in Rational-Comprehensive Decision Theory for Muddling Through

The preceding advances in prescriptive decision theories and methods suggest a number of important implications for the theory and practice of muddling through. Major themes that have emerged since about 1960 include compatibility of incremental adjustment learning techniques with rational-comprehensive optimization; clarification of conditions under which these incremental techniques fail; conditions required for them to succeed; ways to speed and improve them; and new insights into design of incentives, communications, and policy adaptation procedures for teams of cooperating agents and for sets of players with different interests.

Rational-Comprehensive Decision Theory vs. Muddling Through: A False Dichotomy

The rational and comprehensive view of decision-making which Lindblom criticized as unrealistic for many problems that bureaucracies tackle has become substantially more realistic through its vigorous development and adoption of learning-based approaches. RL’s gradual incremental adjustments in response to experience capture a key aspect of muddling through and express it in precise algorithmic formulations that can be analyzed, studied, and improved. A key insight from machine learning is that policy gradient algorithms and other RL and MARL techniques that take successive incremental steps guided by experience – and in this sense muddle through – actually end up solving dynamic optimization problems (e.g., ergodic Markov decision processes (MDPs)). That is, they converge to the same policies as exact optimization methods traditionally used in operations research and statistical decision theory, such as value iteration and policy iteration algorithms in stochastic dynamic programming. This finding addresses the “rational” component of Lindblom’s critique by showing that muddling through and optimization are not opposed. To the contrary, muddling through (as implemented via RL and MARL) provides a constructive way to solve large classes of dynamic optimization problems. RL also addresses the “comprehensive” component of Lindblom’s critique. Its ability to solve a variety of adaptive dynamic optimization

---

68 Pham et al. 2018.
69 Shalev-Schwartz et al. 2018.
70 Cai et al. 2013.
problems by iterative improvement algorithms that do not require initial knowledge of the optimization problems being solved – and specifically of how different choices affect reward probabilities and next-state transition probabilities in dynamic systems or environments – renders the “comprehensive” knowledge requirement no longer necessary. Sampling-based approximate optimization algorithms such as MCTS further reduce the need for a comprehensive examination and evaluation of decision options. In short, rather than being thesis and antithesis, optimization and muddling through have undergone a useful synthesis in modern machine learning via RL and MARL.

**Limitations and Failures of Incremental Learning**

Incremental learning and improvement processes for policy formulation, as modeled by RL and MARL algorithms, can fail in two main ways: by failing to converge (either at all or quickly enough to be useful) or by converging to a non-optimal policy. It is common experience in practice that RL and MARL algorithms do not converge quickly to optimal policies\(^\text{71}\). Even in idealized applications such as learning to play video games, many millions of frames of training data and days to weeks of CPU time are needed to train state-of-the-art RL algorithms to reach or surpass human levels of performance\(^\text{72}\). Moreover, unless reward functions are carefully crafted and RL principles meticulously implemented with well-chosen learning rates, RL may deliver a needlessly poor policy – one in which higher rewards could be achieved using a different feasible policy. This can happen if exploration gets stuck in a local optimum or if the reward functions used in training do not elicit policies that generalize and perform well in practice outside the training data set\(^\text{73}\). Techniques to overcome these challenges are being very actively developed (see Table 2). They include meta-learning, which helps adjust quickly to new situations\(^\text{74}\); “transfer learning” using invariant causal models of how actions affect reward and next-state probabilities that can be generalized or transferred between training and application settings\(^\text{75}\); and imitation learning algorithms that relatively quickly reverse-engineer the observed successes of others\(^\text{76}\). But fully automated RL and MARL techniques for quickly discovering optimal or near-optimal policies remain elusive. Fundamental computational complexity results for decentralized control of Markov decision processes (MDPs) and their generalizations suggest that some of these limitations are intrinsic for MARL (although not for single-agent RL with MDPs)\(^\text{77}\), and hence that discovery of high-reward policies will always be time-consuming unless there is some measure of centralized control\(^\text{78}\).

\(^{71}\) Irpan 2018.


\(^{73}\) Irpan 2018.

\(^{74}\) Clavira et al. 2018.

\(^{75}\) Rojas-Carulla et al. 2018; Zhang and Bareinboim 2017.

\(^{76}\) Shiarlis et al. 2016.

\(^{77}\) Papadimitriou and Tsiakis 1985.

\(^{78}\) Bernstein et al. 2000.
Of course, real organizations do not simply implement computer science algorithms, and it would be simplistic to read into the complexities of human organizational design and behavior all the limitations (or only the limitations) of RL and MARL algorithms. Nonetheless, understanding how and why these algorithms fail in some settings suggests important pitfalls to avoid in organizations that rely on muddling through, insofar as they follow the same basic principles. To the extent that muddling through is held out as being a normatively desirable approach to practical decision-making in an uncertain world, RL algorithms that deploy the same ideas should demonstrate good performance in simple, well-understood settings such as MDPs — and, indeed, they often do\textsuperscript{79}. Conversely, several success factors that turn out to be necessary for effective RL or MARL machine learning of high-reward policies in relatively simple environments may help to suggest necessary (although not sufficient) conditions for effective organizational learning within and among human organizations. To these we turn next.

**Success Factors and Principles for More Effective Muddling Through**

The past two decades of experience with RL, MARL, and game-playing agents have uncovered valuable lessons about success factors and principles for effective incremental learning and policy improvement in uncertain environments. These principles are likely to be important for any normative decision theory of muddling through that provably or demonstrably leads to high-reward policies in complex, uncertain, changing, and large-scale environments, especially when there are delays between choices and consequences. The following paragraphs summarize key lessons and some comparisons with observed real-world decision processes for human organizations.

1. **Collect accurate, relevant feedback data and use it to improve policies.**

After each new action is taken, RL evaluates the reward received and compares it to the reward that was expected so that the difference can be used to correct erroneous expectations and update the current policy. This requires that the effects of actions be evaluated and compared to prior expectations or predictions, and also that policies then be adjusted in light of the data. In the real world, policy-making and policy-administering bureaucracies frequently violate each of these requirements. For example, finding that investments in a costly course of action have yielded lower-than-expected returns may provoke those who originally chose it to escalate their commitment to it\textsuperscript{80}. Where principles of rational decision-making would prescribe ignoring sunk costs and searching for a higher-reward alternative instead, real decision-makers are often reluctant to acknowledge disappointment or failure and may seek to justify their initial decisions to themselves or others, leading to predictable and easily observed departures from rationality. The

---

\textsuperscript{79} Dai et al. 2018.

\textsuperscript{80} Molden and Hui 2011; Schultze et al. 2012.
often-observed tendency of investors to hold onto declining stocks for too long is one example. Possible psychological and political explanations for escalating commitment range from loss aversion to seeking to manage the impressions of others, but clearly such resistance to modifying or abandoning previous choices in light of experience inhibits effective learning. Many other well-studied psychological biases also block effective learning in individuals, groups, and organizations.\footnote{Cox 2015; Tetlock and Gardner 2015.}

Even when the definition of reward is clear, e.g., reductions in mortality rates or illness-days, measurement errors for states and rewards (and for actions, if monitoring of policy implementation is imperfect) can mislead machine learning techniques into recommending dominated policies by distorting estimated reward probabilities for different act-state pairs.\footnote{Mullainathan and Obermeyer 2017.} In business as well as government, data needed to evaluate and compare actual to predicted performance of a policy are often not even collected, or are ignored or misinterpreted if they are collected. A common avoidable decision trap leading to needlessly poor (low-reward) policies has been described as “Fooling Yourself About Feedback—Failing to interpret the evidence from past outcomes for what it really says, either because you are protecting your ego or because you are tricked by hindsight”\footnote{Russo and Schoemaker 1989.}. In social policy application areas as diverse as education, criminal justice, and healthcare, changes in policy are often implemented without any clear predictions about expected changes in rewards or careful evaluations of actual changes in rewards.\footnote{Tetlock and Gardner 2015.} These failures of design and analysis preclude using the differences between predicted and observed values to improve predictive models and policy choices. They prevent the crucial learning from experience that is essential to effective muddling through.

The remedy is to collect, retain, candidly communicate, and use accurate data on predicted and observed outcomes from implemented policies to improve them over time. Simple and sensible as this prescription appears – who would question the wisdom of using relevant feedback data to improve decisions? – it is so widely neglected in current practice that there are many opportunities to improve current policy-making in industry, government at all levels, and international collaborations simply by following it. The following examples illustrate the potential practical relevance, value, and pitfalls of collecting and then using relevant feedback information.

**Example: Accountability Studies in Air Pollution Health Effects Research**

Air pollution health effects research has recently started to use “accountability studies” that examine the real-world consequences over time of policy interventions intended to reduce air pollution levels and adverse health effects.\footnote{Boogaard et al. 2017; Henneman et al. 2017.} A repeated finding is that beneficial effects of
interventions that have been confidently predicted or estimated from statistical exposure-response associations and models are not observed in valid accountability studies. For example, bans on smoky coal in Ireland that successfully reduced particulate air pollution by about 40%-70% and several dozen micrograms per cubic meter in various locations were originally predicted (and claimed) to thereby have substantially reduced total and cardiovascular mortality rates. A typical claim was that “The net benefit of the reduced death rate was greater than predicted from results of previous time-series studies”\textsuperscript{86}. However, subsequent accountability research found that the intervention actually had no detectable effect on causing reduced total or cardiovascular death rates, which declined equally quickly in areas where pollution decreased and where it did not. As explained by other researchers in retrospect, “However, even when studying an abrupt action, threats to causal validity can arise, as illustrated in extended analyses of the Dublin coal ban that revealed that long-term trends in cardiovascular health spanning implementation of the ban — not the coal ban itself—contributed to apparent effects on cardiovascular mortality.”\textsuperscript{87}

Rather than acknowledging and learning from this disappointing experience and seeking more effective policies, the Irish government simply ignored the unwelcome data. As of 2018, years after the bans were well known to have no detectable effect on total death rates, the risk information communicated to policy makers and the public was as follows: “The original ban in Dublin is cited widely as a successful policy measure both nationally and internationally as best practice within the international clean air community. It is estimated that approximately 8,000 premature deaths have been averted in Dublin since the introduction of the smoky coal ban in 1990. Further health, environmental and economic benefits (estimated at €53m per year) are expected, when the ban is extended nationwide”\textsuperscript{88}. In principle, such a wide gap between the substantial benefits in reduced total death rates predicted using statistical modeling and attribution assumptions and the lack of such benefits observed in reality constitutes a strong signal from which to learn. Both scientists and policy makers should be eager to understand why the predicted public health benefits of large reductions in particulate air pollution did not materialize, and to use this understanding to design policies that are more successful in causing desired improvements in life expectancy, perhaps by focusing on locations with more toxic compositions of air pollution. Instead, the propagation of false feedback signals (“It is estimated that approximately 8,000 premature deaths have been averted”) leads to unrealistic expectations and consequent continuation and expansion of policies that have been found not to work (“Further health, environmental and economic benefits (estimated at €53m per year) are expected, when the ban is extended nationwide”). Accountability studies, by implementing the most fundamental principle of incremental learning and improvement – collect data and compare predicted/expected to actual observed outcomes – can help to improve upon current practices that ignore or deny disappointing feedback.

\textsuperscript{86} Clancy et al. 2002. \\
\textsuperscript{87} Zigler and Dominici 2014. \\
\textsuperscript{88} Ireland EPA 2018.
Example: Animal Antibiotics Bans

In 2017, the World Health Organization (WHO) recommended “that farmers and the food industry stop using antibiotics routinely to promote growth and prevent disease in healthy animals” in order “to prevent the spread of antibiotic resistance.” They noted that “interventions that restrict antibiotic use in food-producing animals reduced antibiotic-resistant bacteria in these animals by up to 39%. This research directly informed the development of WHO’s new guidelines.”\(^{89}\). This illustrates use of an irrelevant or questionable feedback signal – reduction of antibiotic-resistant bacteria in healthy food animals – to support recommended policies that have repeatedly been found to be ineffective for achieving stated goals such as preventing or slowing spread of antibiotic-resistant illnesses in human patients. Three decades of experiences in multiple countries have repeatedly found that bans and restrictions of antibiotics used to promote healthy growth of food animals have produced no detectable benefit in reducing or delaying spread of antibiotic resistance in human patients, perhaps because human antibiotic-resistant bacterial infections arise largely from human hospital-adapted strains of bacteria that are unaffected by antibiotic use on farms. Typical findings are that “Neither the Netherlands' nor Denmark's antimicrobial ban has resulted in decreased antimicrobial resistance in humans. In addition, a study performed in the Netherlands concluded that the therapeutic use of antimicrobials in food animals has nearly doubled in the past decade – one of the likely factors in that increase is the ban on the use of antimicrobials for growth promotion.”\(^{90}\)

Although bans and restrictions such as those that WHO recommends have been ineffective in achieving their stated goals, they have greatly increased diseases and painful deaths in food animals, especially necrotic enteritis in poultry and dysentery in swine, thereby significantly reducing the productivity and increasing the costs of food production operations\(^{91}\). By conventional cost-benefit metrics, then, these might seem to be poor policies. However, they remain very politically popular in Europe and the United States, among other places. They are strongly advocated by various consumer and scientific activist groups who maintain that blocking animal antibiotic use is vitally important for slowing spread of antibiotic resistance in humans. Evidence to the contrary, such as that just mentioned, is not simply disregarded by these concerned groups, but is often replaced by focusing on less relevant evidence that seems to better support a pro-restriction agenda, such as that bans on animal antibiotic do reduce resistance in bacteria from healthy animals, even if they don’t reduce risks to human patients. The WHO rationale quoted above is a case in point. This example highlights that even seemingly uncontroversial advice, such as to collect and learn from relevant feedback data, which is straightforward to implement in many RL and MARL applications, can be difficult to implement in human policy making, where evidence is often sought and used to support preferred actions, rather than to learn which actions

\(^{89}\) WHO 2017.
\(^{90}\) AVMA 2018.
\(^{91}\) e.g., von Immerseel et al. 2009, M'Sadeq et al. 2015.
should be preferred for achieving stated goals. More generally, the lessons for effective muddling through drawn from experiences in machine learning may be necessary in human organizations, too, but they are not sufficient, insofar as they ignore many aspects of decision psychology and organizational politics (e.g., impression management) that complicate human learning, cooperation, and decision processes and that have no clear counterparts in normative decision theories for rational agents.

2. Explore via experiments to discover how to cause desired changes in outcome probabilities.

It is tempting for a policy analyst or policy maker steeped in the rational-comprehensive tradition criticized by Lindblom to create the best possible model of how one believes the world works – quantifying the expected utility or reward from different actions or policies in different states, together with probabilities for the different states to capture uncertainty – and then to choose the action or policy that maximized expected utility according to this model, as in equation (1). But in reality, the causal relationship between choices of policies and resulting conditional probabilities of different consequences and rewards is often initially highly uncertain. Prudent and effective policy-making require acknowledging and coping with this *model uncertainty*, rather than selecting and using a single model. RL and MARL algorithms do this in the following ways:

- **Experiment using randomization:** An intrinsic part of many RL algorithms is randomized selection of actions using Thompson sampling or other randomized sampling schemes. Such experimentation helps to discover which policies work best and to avoid becoming stuck in local optima, but it is counter-cultural among people who believe that one should know and not guess about the best course of action before taking it, as well as among decision analysts who believe that one should solve an expected utility optimization problem and then make deterministic decisions based on the results. Neither belief fully acknowledges or responds constructively to the reality emphasized by Lindblom, that current knowledge is often simply insufficient to permit confident identification of the best policy, and that experimentation is the only practical way to discover how to do better. Fortunately, use of randomized controlled trials (RCTs) in social policy experimentation and evaluation of interventions has become increasingly accepted and practiced recently, in areas ranging from disrupting poverty to preventing delinquency to improving oral health of fifth grade students to reducing child abuse by intervening

---

92 Kahneman 2011.
93 Schulze and Evans 2018.
94 Tetlock and Gardner 2015.
95 Tollefson 2015.
96 de Vries et al. 2018.
97 Qadri et al. 2018.
with substance-abusing parents\textsuperscript{98}. For learning to control air pollution health effects, RCTs may not be practicable or ethical, but natural experiments and quasi-experiments provide valuable opportunities to learn from observed responses to unplanned or non-random interventions\textsuperscript{99}.

- **Use causal models to guide safe exploration, reduction of current model uncertainties, and improvement of current policies.** Discovering high-reward policies usually requires model-based search because model-free learning takes too long. Causal graph models can be learned from data, including data collected from small interventions\textsuperscript{100}, and used to predict how the conditional probability distributions of outcome and reward variables will change in response to changes in their direct or indirect causes (i.e., their parents or ancestors in a causal directed acyclic graph (DAG) model), including policy variables controlled by a decision-maker\textsuperscript{101}). For example, data collected for accountability studies of the health effects of the Dublin coal burning ban could be used to discover whether, empirically, mortality rates are conditionally independent of particulate pollution concentrations given the year (as in the causal graph model: pollution concentration $\leftarrow$ year $\rightarrow$ mortality rate), or whether mortality rates vary with particulate pollution concentrations even when year is held fixed (as in the alternative model: year $\rightarrow$ pollution concentration $\rightarrow$ mortality rate). In the second model, but not the first, exogenous changes in pollution concentration are predicted to change mortality rates. Data can be used both to discriminate between these alternative model structures and also to quantify the conditional probabilities for the values of each variable given the values of the variables that point into it\textsuperscript{102}. Such causal models, when they can be uniquely identified from data, enable the effects of interventions on outcomes of interest to be predicted even in novel settings, provided that the conditional probability tables or models in the causal model are invariant across settings – often taken as a defining property of valid causal laws and models\textsuperscript{103}. This invariance of causal relations also provides a principled basis for transfer learning and for generalizing results of experiments or randomized controlled trials to new settings, as well as for synthesizing estimates of causal impacts across multiple studies\textsuperscript{104}. When a causal model cannot be uniquely identified from available data, bounds on causal impacts and ensembles of possible alternative models that are consistent with the data can and should be used to characterize model uncertainty. This allows for safe learning and safe exploration in which small changes in policies are allowed in order to search for better ones, but...
permitted changes are restricted to avoid regions of high model uncertainty that could cause very poor outcomes (Table 3). Finally, adroit use of multiple models and meta-learning can help to speed adaptation to new and unexpected conditions and avoid the pitfalls of relying on a single model that may be incorrect or that may become outdated.\footnote{Clavira et al. 2018.}

3. During collective learning, agents should advance slowly when doing better than expected, but retreat quickly when doing worse.

This restates the “win or lose fast” (WoLF) principle from MARL. “Advance slowly” means use small learning rates and step sizes to adjust evaluations of act-state pairs and to change policies (conditional probabilities of taking different actions in each state, or given available information). “Retreat quickly” means use larger learning rates and step sizes. More generally, it is necessary to carefully control step sizes to avoid destabilizing the learning and improvement process. This advice does not contradict principles for single-person sequential decision-making, such as the optimality of bold play (staking one’s fortune) when trying to reach a target before being ruined and playing against unfavorable odds\footnote{Dubins and Savage, 1965.}; or the principle of optimism in the face of uncertainty for sequential decision-making under uncertainty, e.g., in multi-arm bandit problems\footnote{Munos 2014.}. Rather, WoLF provides a useful heuristic for coordinating the rates at which agents on a team adjust their individual policies to prevent collective instability, so that they can eventually find and exploit a coordinated set of individual policies for maximizing team reward.

In practice, destabilized policy-making processes in human organizations can manifest as “policy churn,” in which new policies are proposed before old ones are well implemented and evaluated by the teams of agents implementing them\footnote{Monios 2016.}. Teachers implementing education reform programs; bankers implementing new risk management regulations and requirements; medical staff implementing new infection control protocols in hospital wards; and workers in bureaucracies implementing policy changes have all been frustrated by policy churn. It leads to costly activity and change without providing the opportunities for careful and thorough evaluation and improvement needed to improve outcomes. Indeed, it is tempting to speculate that perhaps fear of constant deflections and resulting lack of progress might motivate some of the previously discussed reluctance to systematically collect and use feedback data to evaluate and improve policies. Conversely, desire to show action and strong leadership, or to obscure the results of previous ineffective choices, might provide incentives for policy churn. In any case, the study of RL and MARL algorithm performance suggests the optimistic possibility that deliberately controlling step sizes and adjustment rates for policy updates might facilitate productive

\footnote{Clavira et al.
2018.}
\footnote{Dubins and Savage, 1965.}
\footnote{Munos 2014.}
\footnote{Monios 2016.}
incorporation of feedback data into policy updates for a group of cooperating agents without destabilizing their learning and improvement process.

4. Separate actors and critics.

Discussing step sizes and adjustment rates in abstract mathematical terms is one thing; translating the discussion to practice in a workplace is another. In RL algorithms, the learning rate parameter has a clear interpretation as the size of the adjustment in value estimates or policy parameters, such as conditional probabilities for selecting actions given observations, made in response to an observed difference between predicted and received rewards. In the much more complex, constrained, and messy real world of human governance, policy making, and group or organizational decisions, any change in a current policy is usually costly. The RL idealization of frequent small adjustments made without significant costs, delays, or uncertainties in implementation is too simple to describe most real-world decision processes. Nonetheless, some RL and MARL principles may still be useful for human organizations. One of the most useful may be that decision and evaluation of decision performance should be kept distinct processes.

Reasons abound in individual and group psychology for keeping those who make decisions about policy adjustments (analogous to “actors” in actor-critic RL algorithms) separate from those who evaluate the performance of the policies and provide feedback and suggestions for improving them (the “critics”). Among these reasons are confirmation bias, motivated reasoning, groupthink, and other heuristics and biases\textsuperscript{109}. RL suggests an additional reason, rooted in statistics: in deep learning RL algorithms, training one network to decide what to do next and a separate one to evaluate how well it is working has been found to prevent overly optimistic assessments of policy performance due to overfitting, i.e., using the same data to both select estimated value-maximizing actions and estimate the values from taking those actions. Preventing the statistical over-optimism that arises when the same network is used to select policies and to evaluate them can significantly improve the speed and quality of the learning and improvement process in discovering truly high-reward policies\textsuperscript{110}. The principle of separating the processes for choosing which changes to make and evaluating how well they perform can also be applied usefully to choice of learning rates (i.e., choosing how quickly to modify current policies in light of feedback) as well as to choice of policies\textsuperscript{111}. Possible future advances include deliberately diversifying the learning rates of different agents on the same team to obtain the advantages of both rapid exploration of new policies and thorough exploitation and refinement of old ones. This is an old concept in organizational science\textsuperscript{112}, but is still being developed in MARL research, largely through sub-literatures on

\textsuperscript{109} Cox 2015.
\textsuperscript{110} van Hesselt et al. 2015.
\textsuperscript{111} Xu et al. 2017.
\textsuperscript{112} e.g. March, 1991.
heterogeneous learning agents\textsuperscript{113}, parallelization of search, and extensions of the previously discussed WoLF (win or lose fast) principle.

As a practical matter, separation of actors and critics can be applied fruitfully to major social learning and improvement initiatives, such as air pollution regulation, through accountability studies that revisit previous regulatory actions or other decisions to assess the results they have produced\textsuperscript{114}. Greater use of such evaluation studies to evaluate and update previous policy decisions – ideally, in time to be useful in guiding policy decisions elsewhere – is clearly consistent with the principle of collecting and using relevant feedback data. Separation of actors and critics provides an additional principle for using feedback data to maximum advantage to improve polices and their results.

5. Shape rewards to promote learning and improvement.

How individual agents are rewarded in response to team accomplishment creates incentives affecting their choices of what actions to take under uncertainty and what to communicate to others in an organization, as mechanism design and team theory have greatly clarified. Rewards allocated to individual agents can also play an important additional role by providing quantitative feedback about how much each agent is contributing to a team’s progress in increasing collective reward obtained from the environment, thus yielding useful information as well as incentives to help each agent learn to behave in more valuable ways. The notorious “credit assignment problem” of estimating how much each agent’s choices contributed to a team’s eventual outcomes (e.g., winning or losing a chess match, scoring a goal or being scored on in robot soccer, or adding to or reducing profitability or sustainability or share price in a business) is a challenging problem in RL and MARL.

Recently, it has been found that using causal (counterfactual) models to shape each agent’s reward to reflect the estimated difference it has made – the difference between what was actually achieved and what would have been expected without each agent’s contribution, or its marginal value, in microeconomic terms – can speed collective learning and optimization when each agent seeks to maximize its own reward\textsuperscript{115}. This research uses mathematical rewards that are costless to implement, so that budget constraints such as that the sum of agent rewards must not exceed the collective reward of the team, do not apply. However, it seems plausible that, even in the presence of budget constraints, rewarding each agent according to its estimated marginal contribution (or its expected marginal contributions, for Shapley values in non-cooperative game theory) might promote joint learning about how to contribute more effectively, as well as having other properties of efficiency and fairness familiar from microeconomics and game theory. Of course, the

\textsuperscript{113} Potter et al. 2001.
\textsuperscript{114} Boogaard et al. 2017; Henneman et al. 2017.
\textsuperscript{115} Devlin et al. 2014.
asymmetric information about relative roles of chance and effort typical in principal-agent problems can inhibit accurate reward-shaping in practice, and causal modeling of individual marginal contributions to team performance in human organizations is no easy task. Nonetheless, research on how best to use reward shaping to provide feedback and encourage effective learning, as well as to create incentives, may be useful for human organizations as well as for MARL algorithms.

Organizational design and management science literatures are replete with advice on how to shape financial compensation and non-financial rewards and incentives for individual workers and for teams to elicit desired behaviors. Well-developed theories in economics and game theory have characterized what can and cannot be achieved via incentives in the presence of private and asymmetric information\textsuperscript{116}. However, following relatively simple principles, such as rewarding achievement of desired goals rather than easier-to-achieve substitutes for them, can probably do considerable good. For example, air pollution reduction initiatives designed to save lives should be rewarded, praised and promulgated only to the extent that they actually do so. In much of the real world of cooperative government and university research, many prizes are awarded, research proposals funded, press releases issued, and careers advanced for proposing or completing well-intended innovations (e.g., a successful ban on sales of smoky coal or animal antibiotics) even if they do not cause their intended desired consequences, such as demonstrably improving population health and longevity. Conversely, merely discovering and doing what works to produce desired consequences may attract relatively little recognition and reward, especially if it is costly and unglamorous – e.g., reducing spread of antibiotic-resistant infections in human patients by stringent controls on hospital hygiene and antibiotic use. Truisms such as “You get what you pay for,” or “Humans respond to incentives,” remain potent guides to improving current practices and outcomes. Reshaping rewards to emphasize achievement of demonstrated successful results over both well-articulated good intentions and achievement of substitute outcomes (such as reducing coal sales rather than mortality rates, or resistance in bacterial isolates from healthy animals rather than illness-days in human patients) might help to speed more effective experimentation and learning of how to achieve desired results.

It would be naïve to expect such reorientation of rewards to be easy, not least because of the credit assignment problem for sparse and delayed results, and the need for usefully accurate causal models to estimate how the actions taken affect probabilities of valued outcomes. But perhaps solutions emerging from MARL research, including using causal models for credit assignment and timely reward of actions that contribute to causing valued but delayed outcomes, can also help to better shape incentives and rewards in human organizations to facilitate learning and improvement.

\textsuperscript{116} Miceli 2017.
6. Learn from the experiences and expertise of others.

Learning from each other enables agents to speed collective learning. In MARL applications, it helps teams of agents discover more quickly high-reward joint policies for controlling large-scale systems and accomplishing tasks in complex, changing, uncertain environments. We have mentioned several technical tactics for doing so in MARL by having agents share with each other valuable memories, experiences, and expertise (typically encoded as causal models or trained neural nets). Useful techniques include the following:

- **Imitation learning**, where agents observe the successes and failures of others working on the same or similar tasks and reverse-engineer their own successful policies from these observations\(^\text{117}\);
- **Decentralized cooperative communication**, where agents share relevant experiences and data with each other (e.g., local state-action-reward-next-state data or policy and reward data);
- **Experience replay**, i.e., enriching current experiences with recollections of past experiences\(^\text{118}\);
- **Hierarchical communication and coordination.** Studies of hierarchical MARL\(^\text{119}\) indicate that a coordinating center that collects and shares such information can greatly speed learning of high-reward policies, as can decentralized communication among agents working on similar or related tasks, perhaps facilitated by local coordinating supervisors\(^\text{120}\). Recent experiments suggest that even a two-level communication and coordination hierarchy enables teams of MARL agents to control some large-scale and spatially distributed systems, such as supply chains or networks, extremely effectively compared to humans\(^\text{121}\).
- **Transfer learning**, in which the policies learned by some agents (represented by deep neural networks for mapping observations to actions) are transferred to and modified by others, have also been effective in experiments in distributed control of supply chains\(^\text{122}\).
- **Model-based adaptation.** If recent experiences may suddenly become obsolete when changes occur in a controlled system or its environment, it becomes valuable to be able to quickly adapt to new conditions described by different models. These may have been learned previously under different conditions, as in meta-learning\(^\text{123}\); or they may be

\(^{117}\) Shiaričs et al. 2016.
\(^{118}\) Andrychowicz et al. 2018.
\(^{120}\) Zhang et al. 2008.
\(^{121}\) Fuji et al. 2018.
\(^{122}\) Oroojlooyjadid et al. 2018.
\(^{123}\) Lemke et al. 2015; Cai et al. 2018.
developed by generalizing from previous experiences using causal models (i.e., causal transfer learning\textsuperscript{124}).

In applying such ideas to human organizations, it is valuable to recognize that the “agents” may themselves be organizations, such as different schools, hospitals, or companies; or similar government bureaucracies in different states or countries. The United States EPA might learn from the experience of the Irish EPA and start developing and validating causal models to predict where reducing particulate air pollution will have the most and least benefits for human health. States and counties implementing pollution-reducing regulations might learn from each others’ experiences about which combinations of interventions and conditions (possibly involving copollutants, chemical composition of particulate matter, comorbidities and sociodemographic characteristics of the exposed population, weather variables such as temperature and humidity, and so forth) generate the greatest public health benefits from pollution reduction. Failing to learn from the experiences of other jurisdictions and agencies wastes potentially valuable information and needlessly prolongs the time to identify policies that create the largest achievable benefits.

As usual, effective learning in human organizations must overcome challenges from various types of learning aversion that have no clear counterparts in machine learning\textsuperscript{125}. For example, human bureaucracies may reorganize to visibly mimic organizational structures in more successful organizations whose reputations they covet, but without corresponding learning of the more effective policies that drive improved performance\textsuperscript{126}. Conspicuous imitation of form without true imitation learning of competence can help to boost market share even if it does not improve collective learning in ways that increase rewards obtained from the environment. More generally, the possibility that human agents, whether individuals or organizations, will adapt primarily to each other rather than to their environment, thereby failing to capture Pareto-efficient outcomes, is a constant threat in many games with different rewards for different players. Players preoccupied with managing the perceptions and impressions of others to shape allocations of collective efforts and rewards to their own individual advantages may be unable to achieve Pareto efficiency or to maximize any measure of collective success or reward, as demonstrated by numerous impossibility theorems in collective choice theory, game theory, and micro-economic models of incentives and performance in principal-agent relationships, contracts with asymmetric information, and more elaborate organizational structures\textsuperscript{127}. These threats do not arise for teams of agents trying to cooperate in maximizing the same reward function. Our recommendation that agents should learn from each other in order to speed mastery of joint policies for obtaining high rewards from the environment is primarily applicable to such teams of cooperating agents.

\textsuperscript{124} Rojas-Carulla et al. 2018; Zhang and Bareinboim 2017.
\textsuperscript{125} Cox 2015.
\textsuperscript{126} Monios 2016.
\textsuperscript{127} See Table 2 and Cox 2015.
Conclusions

In 1973, two professors of design and city planning at the University of California at Berkeley offered the following sober assessment of the prospects for scientifically based social policy:

“The search for scientific bases for confronting problems of social policy is bound to fail, because of the nature of these problems. They are ‘wicked’ problems, whereas science has developed to deal with ‘tame’ problems. Policy problems cannot be definitively described. Moreover, in a pluralistic society there is nothing like the undisputable public good; there is no objective definition of equity; policies that respond to social problems cannot be meaningfully correct or false; and it makes no sense to talk about ‘optimal solutions’ to social problems unless severe qualifications are imposed first. Even worse, there are no ‘solutions’ in the sense of definitive and objective answers.”

We believe that subsequent developments warrant greater optimism. While it is true that sufficiently heterogeneous or incoherent preferences may make it impracticable, or even impossible, to define and measure a single indisputable public good to be optimized, as attested to by the impossibility results in Table 2, it is also true that agents with at least some shared goals have already achieved impressive feats of cooperation and control in applications as diverse as drone swarm and autonomous vehicle fleet control, search-and-rescue via teams of cooperating autonomous robots, improved management of supply chains, and military gaming using principles of MARL. Such applications are admittedly far less difficult than the wicked problems referred to by Rittel and Webber, but many of the differences are of scale rather than of kind: drone swarms and search-and-rescue robot teams and robot soccer teams are already confronting, with increasing competence, the difficulties of distributed decision-making with initially unclear roles and priorities, uncertain and changing environments, opportunistic revision of goals and plans, and local information that may be time consuming and expensive to share. Muddling through, in the form of RL, MARL, and HMARL algorithms, plays a fundamental role in many of the most successful algorithms for managing risks and uncertainties and successfully achieving goals or completing missions under uncertainty. It is often more practicable to improve via small steps than to try to optimize in one big decision, and this insight from Lindblom’s 1959 paper remains true for machine learning as well human organizations. It has been augmented by the discovery that successive incremental improvement based on feedback at each step and careful selection of step sizes is often an effective way to solve dynamic optimization problems when they can be clearly formulated, as well as an effective way to learn how to act when not enough is initially known to formulate a clear decision optimization problem.

As artificial intelligence and machine learning algorithms are tested and improved on increasingly challenging tasks, principles for learning how to manage risks and act effectively in a variety of centralized, decentralized, and hierarchical organizational structures have begun to emerge. We have discussed several based on recent work that uses deep neural networks to approximate value functions used in RL, MARL, and HMARL algorithms. But an optimistic – and perhaps realistic – view is that these principles are only the beginning of what may soon become a substantial flow from multi agent machine learning to human management science of useful principles for improving organizational design and performance, based on discovering what works best for teams of machine learning agents in solving increasingly important, realistic, and difficult real-world risk management problems. These principles will doubtless require modifications and extensions for the human world, since human psychology for both individuals and groups differs greatly from RL and MARL agent programming. But the pace of discovery and progress in using machine learning to solve increasingly large, difficult, and important real-world problems of decision-making under uncertainty is now extremely rapid. Discovering how groups and teams of agents can organize, learn, decide, and adapt more effectively is becoming an experimental and applied science, as well as a theoretical one, in current artificial intelligence and machine learning. It seems likely that this research will produce insights and principles to help tame currently wicked problems that have high stakes for humans.
Acknowledgements

I thank Vicki Bier for valuable comments on an early draft that improved the framing, content and exposition in this paper. Tables 1-4 resulted from that discussion. I thank Susan Dudley for inviting me to prepare a paper and presentation for panel discussion at the Society for Risk Analysis (SRA) 2018 Annual Meeting, with partial support from the Searle Foundation. This paper reflects the enthusiastic comments and suggestions of organizers and participants in the SRA session. I am grateful for the opportunity to share, discuss, and improve these ideas.
References


environments.” In Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2013 IEEE Symposium. 52-59. IEEE.


Shalev-Shwartz, Shai, Shaked Shammah, Amnon Shashua. 2016. “Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving.” https://arxiv.org/pdf/1610.03295.pdf.


