to billions of parameters, or million times more than many of the networks discussed in this section (Ouyang et al. 2022). One way to characterize this power is that such scale solves the problem of variable binding, or dynamic inferencing, that has limited the generality of smaller networks. For example, if trained with sentences of type 1 composed of words of type A, and sentences of type 2 composed of words of type B, such networks would not generalize to 1-sentences with B-words, and 2-sentences with B-words. Large language models perform such generalization routinely, if they are large enough: For instance, they can write technical instructions in the style of Shakespeare, never seen together in the training corpus.

Interestingly, a large scale is necessary for this ability to emerge. Transformers are based on attention, i.e. discovering useful relationships between input tokens. While the performance of large language and image models is not yet fully understood, it is possible that with a large enough scale, such models start learning relationships between abstractions as well. It would be interesting to see if scale has a similar effect in generating complex, robust, multimodal behavior. It may be possible to use existing pre-trained foundation models in language or vision as a starting point, and evolve behavior generation as a modification or augmentation to them. Or perhaps it will be possible to construct a foundation model for behavior from scratch through the imitation of massive datasets? Or maybe neuroevolution methods can be scaled to large models, and behavior discovered through massive simulations? Research on such scaleup forms a most interesting direction for future work.

6.2 Decision Making

Intelligent behavior, as discussed above, focuses on agents that are embedded in a real or simulated physical environment and interact with it through physical sensors and effectors. In contrast, intelligent decision making focuses on behavior strategies that are more abstract and conceptual, such as those in business and society. Neuroevolution can play a large role in decision making as well, but the approaches and opportunities are distinctly different. They often need to take advantage of surrogate modeling, and take advantage of human expertise, as discussed in this section.

6.2.1 Successes and challenges

To begin, note that human organizations today have vast amounts of data that describe their operation: Businesses record interactions with their customers, measure effectiveness of their marketing campaigns, track performance of their supply chains; health-care organizations follow the behavior of patients, measure effectiveness of treatments, track performance of providers; government organizations track crime, spending, health, construction, economy, etc. Such data has made it possible to predict future trends. Predictions are then used to decide on policies, i.e. decision strategies, i.e. prescriptions, in order to maximize performance and minimize cost.

Discovering optimal decision strategies is an excellent opportunity for neuroevolution. Optimal policies are not known; they involve a large number of variables that interact nonlinearly; the observations and outcomes are often partially observable and noisy; often several conflicting objectives, such as performance and cost, must be optimized at the same time. They are therefore well suited for representation in neural networks, and discovery through evolution.

However, a major challenge is that the search for optimal strategies usually cannot be done in the real world itself. Discovery requires exploration, and it is usually unacceptable to explore novel medical treatments with actual patients, or novel investment strategies with actual money. In discovering intelligent behaviors, such exploration is done in simulation, but it is usually not possible to simulate human behavior, biology, or society in sufficient detail.

However, the vast amount of data, and the predictive models that can be built based on them, provide a possible solution: It may be possible to construct data-based surrogate models of the decision-making environment. These models are phenomenological, i.e. they model the statistical correlations of contexts, actions, and outcomes, and do not simulate the actual underlying processes. However, it turns out that understanding these processes is not even necessary: Phenomenological surrogate models are enough to evaluate the decision strategies, and therefore discover good strategies through neuroevolution.

A surprising synergy emerges in this process. If the predictive models are learned at the same time as the decision strategies based on them, they provide a regularization effect, and a curricular learning effect. As a result, the strategies are more robust and easier to learn. This effect will be discussed in the next subsection.

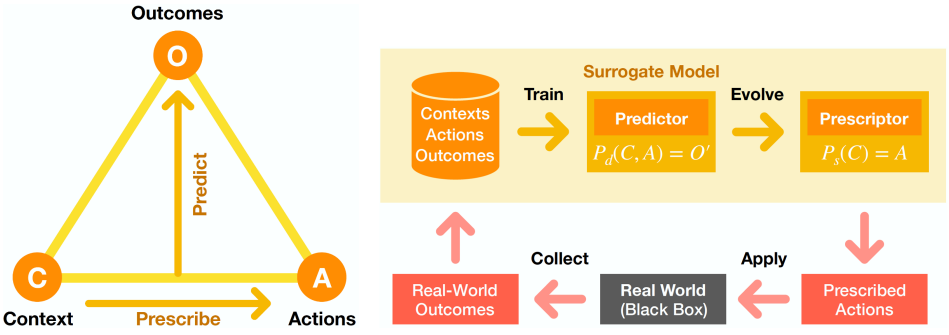
A second challenge in optimizing decision making is that the discovered strategies need to be acceptable to human decision makers. Humans are eventually responsible for deploying them, and in order to do so, they need to be confident that they are indeed good strategies. The strategies need to be trustworthy, i.e. express confidence; they need to make explainable decisions; and it must be possible for the decision makers to interact with them, try out counterfactual scenarios, and convince themselves that the strategies are robust. Considerable work goes into these aspects beyond just neuroevolution of good strategies (Qiu, Meyerson, and Miikkulainen 2020; Miikkulainen, Francon, et al. 2021; Shahrzad, Hodjat, and Miikkulainen 2024).

Part of this challenge is also that there is already significant human expertise in many decision-making domains, and it should be possible to use it as a starting point in discovering better policies. Evolution can still explore, but its exploration is more informed, and may be more likely to discover improvements—also those improvements may be easier for the decision makers to accept. Again it turns out that there is a surprising synergy of human expertise and evolutionary discovery: When put together in this manner, the results are better than either one alone. This effect will be discussed in the second subsection below.

6.2.2 Surrogate modeling

The general idea of discovering decision strategies through surrogate modeling, i.e. the Evolutionary Surrogate-assisted Prescription approach (ESP; not to be confused with the enforced subpopulations method of Sections 5.6 and 7.1.1) is depicted in (Figure 6.9; Francon et al. 2020). The decision-making problem is formalized as a mapping from contexts C and actions A to outcomes O. The goal is to discover a decision strategy, i.e. a prescription policy, that results in the best outcomes for each possible patient.

The starting point is a database, obtained through historical observation, that includes as many examples of this mapping as possible. For instance, C might describe patient characteristics, A might describe procedures or medication, and O might measure the extent and



(*a*) Predictor and prescriptor models (*b*) Surrogate modeling process

Figure 6.9: **Evolutionary surrogate-assisted prescription.** In domains where evaluation of decision strategies is not possible, a surrogate model can be used to guide the search. (*a*) The surrogate model, or a predictor, maps contexts and actions to outcomes. The decision-maker model, or a prescriptor, maps contexts to optimal actions. (*b*) The models are constructed in one or more cycles of an iterative process. Starting from historical observations of contexts, actions, and outcomes, the predictor (e.g. a neural network or a random forest) is trained through supervised learning. It is then used to evaluate prescriptor candidates, constructed through neuroevolution. The final prescriptor is deployed in the domain. More data can then be collected and the cycle repeated, resulting in more accurate predictors and more effective prescriptors. (Figures from Francon et al. 2020)

speed of recovery. This data can be used to train a model, such as a neural network or a random forest, to predict the outcome of a given action in a given context. Thus, the predictor is defined as

$$P_d(C,A) = O', \tag{6.27}$$

such that $\sum_j L(O_j, O'_j)$ across all dimensions j of O is minimized, where L is any of the standard loss functions.

The predictive model in turn can serve as a surrogate in search for good decision strategies. The strategies are mappings themselves, i.e. from contexts to actions, and in particular to actions that result in the best possible outcomes. They are therefore naturally represented as neural networks, and called prescriptive models. The prescriptor takes a given context as input, and outputs a set of actions:

$$P_s(C) = A\,, \tag{6.28}$$

such that $\sum_{i,j} O'_j(C_i, A_i)$ over all possible contexts i is maximized. It thus approximates the optimal decision policy for the problem. Because optimal strategies are not known ahead of time, these models need to be constructed through search, i.e. through neuroevolution. Each candidate is evaluated against the predictor instead of the real world, thus making it possible to explore fully and evaluate a very large number of candidates efficiently.

Once a good candidate is found, it can be deployed in the real world. At this point, uncertainty metrics can be applied to it, it can be distilled into a set of explainable rules, and an interactive scratchpad can be built so that the decision maker can convince him/herself that the policy works as well as expected (Miikkulainen, Francon, et al. 2021). When it is

deployed, more (C, A, O) data can be collected and added to the database. These data are now closer to the actual implemented policies, and make it possible to learn a model that is more accurate where accuracy is most needed. The cycle can then be repeated, resulting in more accurate predictors and more powerful prescriptors in the process.

A practical example of discovering decision strategies for pandemic interventions will be presented in the next subsection. However, in order to evaluate the power of the approach wrt. the state of the art, and to gain insight into how it constructs solutions, it can be implemented in standard reinforcement learning domains (Francon et al. 2020). One good such domain is OpenAI Gym CartPole-v0, i.e. balancing a vertical pendulum by moving a cart left or right. In this case, the process starts with a population of random prescriptors; the predictors are trained at the same time as the prescriptors are evolved, i.e. the loop in Figure 6.9*b* is traversed rapidly many times.

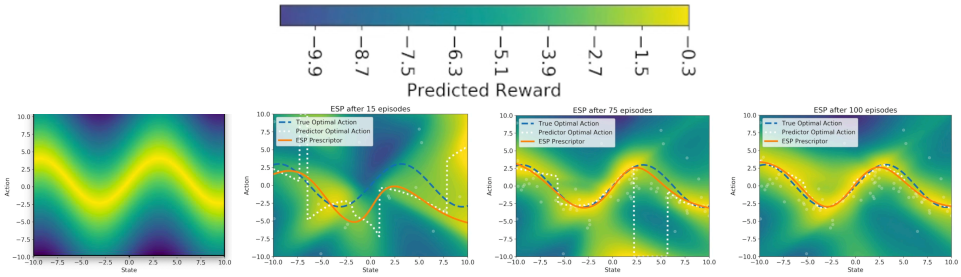
Compared to direct evolution of the control policy as well as standard reinforcement learning methods PPO and DQN, ESP learned significantly faster, found better solutions, had lower variance during search, and lower regret overall. Most importantly, because it is based on the surrogate, ESP is highly sample-efficient, i.e. it requires very few evaluations in the actual domain. Sample efficiency is one of the main challenges in deploying reinforcement learning systems in the real world, and therefore ESP provides a practical alternative.

Such domains are also useful in illustrating how ESP finds solutions. It turns out that they are based on two surprising synergies with learning the predictors. The first one is that such co-learning results in automatic regularization. This effect can be seen most clearly in the domain of evolving function approximators (Figure 6.10). In this case, the context is a scalar value in the x-axis, and the action is a scalar value in the y-axis. The optimal policy is a sine wave; the rewards decrease linearly away from it.

The ESP process starts with randomized feedforward predictor and prescriptor neural networks. In each training episode, a context-action pair is chosen randomly, and the predictor is trained for 2000 epochs with the pairs so far. A population of prescriptors is then evolved for 20 generations, using the same pairs to evaluate them against the current predictor. The top prescriptor is then evaluated against the ground truth to illustrate progress at each episode.

As seen in Figures 6.10*b-f*, after 15 episodes the predictor is still far from representing the sine wave, and the policy optimal wrt. this predictor is highly irregular as well. Remarkably, however, the policy represented by the top prescriptor is much closer to the actual optimal policy. This trend continues throughout training and evolution. By 75 episodes, the top prescriptor has already converged to the optimal policy even though the predictor still suggests an irregular policy, and by 100 episodes, even the predictor-optimal policy is a sine wave. This convergence is remarkably rapid: PPO takes over 3000 episodes to learn a good approximation, and direct evolution (with the predictor) is not even close at that point.

How is it possible for ESP to discover an optimal policy when the predictor is still far from it? It turns out that the simultaneous learning of the predictor provides a regularization effect. The best predictors stay in the population for several generations, and therefore are evaluated against many different versions of the predictors. Especially early on in predictor training, the predictors vary significantly. In a sense, they form an ensemble, and the prescriptors are evaluated against this ensemble. The ensemble performs better than



(*a*) Problem space (*b*) After 15 samples (*c*) After 75 samples (*d*) After 100 samples

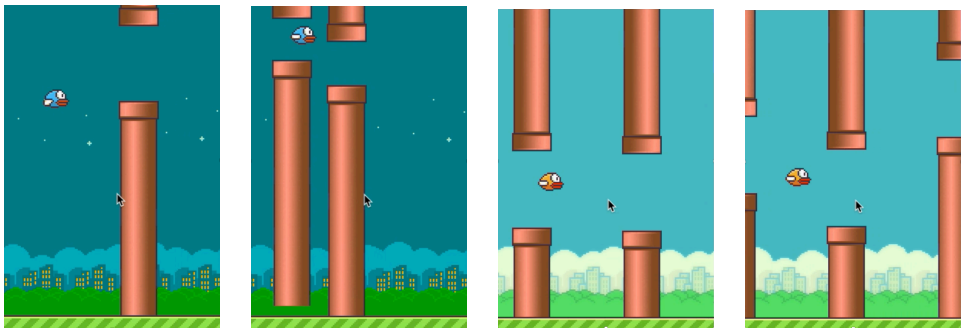
Figure 6.10: **Evolving effective decision making through co-learning of the surrogate model.** This example illustrates the synergy of learning the predictor and prescriptor at the same time in the function approximation domain. (*a*) With the context as x and the action as y, the ground truth outcomes are indicated as the colored background. (*b-d*) The current predictor is indicated similarly as the colored background so that it can be compared with the ground truth in (*a*). The training pairs are illustrated with translucent dots. The actual optimal policy is indicated by the blue dotted line, and the policy that is optimal wrt. the current predictor is shown in white dotted line. The policy represented by the current top predictor is indicated by the solid orange line. The prescriptors evolve policies that are better than the predictors suggest. The prescriptors are evaluated with several different predictors over time, which act as an ensemble that is more accurate than any single predictor alone. Such co-learning of the predictor and the prescriptors thus results in automatic regularization, leading to faster learning and more robust solutions. For an animation of this process see https://neuroevolutionbook.com/neuroevolution-demos. (Figures from Francon et al. 2020)

any individual predictor, and therefore the prescriptor evaluation is more accurate as well. Thus, the co-learning of predictors and prescriptors provides a surprising regularization effect that makes it possible to progress faster than expected.

Another useful effect of co-learning is the curricular learning environment it provides. That is, the early predictors capture the main trends and the most general aspects of the environment, which then become refined as they learn more. Thus, the challenges start simple and become more complex as the training goes on—this is the main principle of curricular learning in general, and a good way to construct complex behavior (as also seen in Section 3.4).

The effect can be made concrete in the Flappy Bird game environment. The bird flies at a constant speed through a series of gates in pipes. The player has only one action, flap, which lifts the bird up a constant amount. Gravity will then bring it rapidly down. The challenge is to time the flaps so that the bird gets through the next gate, and is also well positioned to get through the next gate. In the ESP setup, the predictor is trained to estimate the next game states given the current state and the action, and prescriptors evolved to decide when to flap. The fitness is increased for every gate that the bird successfully clears.

Figure 6.11 shows four sample predictions during evolution. Curricular learning is evident in these snapshots: At the beginning, the predictor tends to place the gate near the bird, making it easy to fly to it. By the time the bird evolves to fly through one gate, the predictor



(*a*) First gate (*b*) Pair of gates (*c*) Straight run (*d*) Full problem

Figure 6.11: **Automatic curricular evolution through co-learning of the surrogate model.** In the FlappyBird game, the challenge is to flap the bird up at the appropriate times so that it flies through a course of gates without hitting them. The predictor, trained to estimate the result of an action (flap/no-flap) at a state, (*a*) first places the gate nearby, (*b*) then clusters a number of them together, (*c*) then spreads them apart at the same level, and (*d*) finally presents the full game challenge accurately. Such a series of increasingly challenging evaluations provides a curriculum that makes it possible to evolve successful behavior, even when it would not evolve with the full challenge from scratch. Co-learning the predictor and prescriptor thus constructs an effective curriculum automatically, allowing neuroevolution to solve more difficult tasks. For animations of these behaviors, see https://neuroevolutionbook.com/neuroevolution-demos.

has learned to expect the next gate, but clusters it together with the first one. It is thus relatively easy to evolve behavior that clears several gates. As the predictor learns, it spreads the gates further apart, but still keeps them roughly at the same level. While the prescriptors evolve to fly straight through, the predictors start placing the gates further up and down, eventually providing a realistic challenge. By that time, it is relatively easy to evolve behavior that takes the height of the gates into account, and flap the bird successfully through the course. In contrast, direct evolution, i.e. evolution from scratch in the actual task, never constructs successful behavior. This result demonstrates the power of curricular learning, and shows how it can be automatically discovered by learning the challenges at the same time as the solutions.

6.2.3 Case study: Mitigating climate change through optimized land use

A significant factor contributing to climate change is how much land area is allocated for different uses (Friedlingstein et al. 2023). Forests in general remove more carbon from the atmosphere than e.g. crops and ranges, yet such uses are essential for the economy. Land-use patterns must therefore be planned to minimize carbon emissions and maximize carbon removal while maintaining economic viability.

An approach to optimize land use can be developed based on the ESP method discussed in the previous section (Miikkulainen, Francon, et al. 2023). The idea is to first utilize historical data to learn a surrogate model on how land-use decisions in different contexts affect carbon emissions and removals. Then, this model is used to evaluate candidates in an evolutionary search process for good land-use change policies. While it is difficult to

predict economic impact of changes in land use, the amount of change can be used as a proxy for it. As a result, a Pareto front is generated of solutions that trade off reduction in carbon emissions and the amount of change in land use. Each point in the Pareto front represents an optimal policy for that tradeoff.

The data for carbon emissions (Emissions resulting from Land-Use Change, ELUC) originate from a high-fidelity simulator called Bookkeeping of Land-Use Emissions (BLUE; (Hansis, Davis, and Pongratz 2015)). BLUE is designed to estimate the long-term CO2 impact of committed land use. "Committed emissions" means all the emissions that are caused by a land-use change event are attributed to the year of the event. BLUE is a bookkeeping model that attributes carbon fluxes to land-use activities. While in principle a simulator can be used as the surrogate model for ESP, in practice the simulations are too expensive to carry out on demand during the search for good policies. Therefore, the BLUE team performed a number of simulations covering a comprehensive set of situations for 1850-2022, resulting in a dataset that could be used to train an efficient surrogate model.

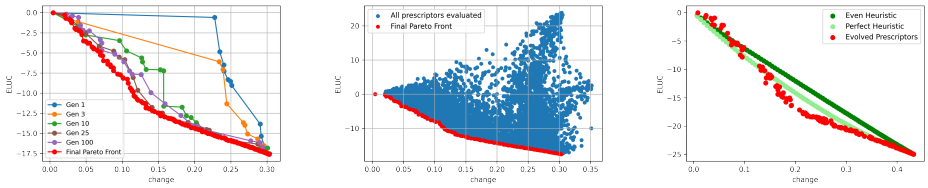
The Land-Use Change (LUC) data is provided by the Land-Use Harmonization project ((LUH2; Hurtt et al. 2020). A land-use harmonization strategy estimates the fractional land-use patterns, underlying land-use transitions, and key agricultural management information, annually for the time period 850-2100 at 0.25 x 0.25 degree resolution.

Based on these data, the modeling approach aims to understand the domain in two ways: (1) In a particular situation, what are the outcomes of the decision maker's actions? (2) What are the decisions that result in the best outcomes, i.e. the lowest carbon emission and cost for each tradeoff between them? The data is thus organized into context, action, and outcome variables.

Context describes the problem the decision maker is facing, i.e. a particular grid cell, a point in time when the decision has to be made, and the usage of the land at that point. More specifically it consists of latitude and longitude and the area of the grid cell, the year, and the percentage of land used in each LUH2 category (as well as nonland, i.e. sea, lake, etc.).

Actions represent the choices the decision-maker faces. How can they change the land? In the study of this paper, these decisions are limited in two ways: First, decision-makers cannot affect primary land. The idea is that it is always better to preserve primary vegetation; destroying it is not an option given to the system. Technically, it is not possible to re-plant primary vegetation. Once destroyed, it is destroyed forever. If re-planted, it would become secondary vegetation. Second, decision-makers cannot affect urban areas. The needs of urban areas are dictated by other imperatives, and optimized by other decision makers. Therefore, the system cannot recommend that a city should be destroyed, or expanded.

Outcomes consist of two conflicting variables. The primary variable is ELUC, i.e. emissions from land-use change. It consists of all CO2 emissions attributed to the change, in metric tons of carbon per hectare (tC/ha), obtained from the BLUE simulation. A positive number means carbon is emitted, a negative number means carbon is captured. The secondary variable is the cost of the change, represented by the percentage of land that was changed. This variable is calculated directly from the actions. There is a trade-off between these two objectives: It is easy to reduce emissions by changing most of the land, but that would come at a huge cost. Therefore, decision-makers have to minimize ELUC while



(*a*) Evolution of Pareto front (*b*) All prescriptors evaluated (*c*) Comparing to heuristics

Figure 6.12: Prescriptor evolution and performance. In the land-use optimization domain, the goal is achieve low carbon emissions with minimal change in land-use. (*a*) The Pareto front moves towards the lower left corner over evolution, finding better implementations for the different tradeoffs of the ELUC and change objectives. (*b*) Each prescriptor evaluated during evolution is shown as a dot, demonstrating a wide variety of solutions and tradeoffs. The final Pareto front is shown as red dots in both figures, constituting a set of solutions from which the decision-maker can choose a preferred one. (*c*) The Pareto fronts of evolved prescriptors vs. heuristic baselines. Whereas the heuristics try to optimize each region equally, the evolved prescriptors allocate more change to where it matters the most. This result demonstrates that the approach can discover non-obvious opportunities in the domain, and thus find better solutions than the obvious heuristics. For an interactive demo of the system, see https://neuroevolutionbook.com/neuroevolution-demos. (Figures from Miikkulainen, Francon, et al. 2023).

minimizing land change at the same time. Consequently, the result is not a single recommendation, but a Pareto front where each point represents the best implementation of each tradeoff given a balance between the two outcomes.

The ESP implementation consists of the predictor, trained with supervised learning on the historical data, and the prescriptor, trained through evolution. Given the context and actions that were performed, the predictive model estimates the outcomes. In this case, since the cost outcome can be calculated directly, only the ELUC is predicted by the model. That is, given the land usage of a specific location, and the changes that were made during a specific year, the model predicts the CO2 long-term emissions directly caused by these changes. Any predictive model can be used in this task, including a neural network, random forest, or linear regression. As usual, the model is fit to the existing historical data and evaluated with left-out data.

Given context, the prescriptive model suggests actions that optimize the outcomes. The model has to do this for all possible contexts, and therefore it represents an entire strategy for optimal land use. The strategy can be implemented in various ways, including decision trees, sets of rules, or neural networks. The current approach is based on neural networks. The optimal actions are not known, but the performance of each candidate strategy can be measured (using the predictive model), therefore the prescriptive model needs to be learned using search techniques such as neuroevolution. As in prior applications of ESP (Francon et al. 2020; Miikkulainen, Francon, et al. 2021), the prescription network has a fixed architecture of two fully connected layers; its weights are concatenated into a vector and evolved through crossover and mutation.

In preliminary experiments, prediction performance was found to differ between major geographical regions. To make these differences explicit, separate models were trained on different subsets of countries: Western Europe (EU), South America (SA), and the United States (US). Three different predictive models were evaluated: linear regression (LinReg), Random Forests (RF), and neural networks (NeuralNet). They were trained with a sampling of data upto 2011, and were tested with data from [2012-2021]. Not surprisingly, in each region the models trained on that region performed the best. The LinReg models performed consistently the worst, suggesting that the problem includes significant nonlinear dependencies. RF performed significantly better; however, RF does not extrapolate well beyond the training examples. In contrast, neural nets both capture nonlinearities and extrapolate well, and turned out to be best models overall. Therefore, the global neural net surrogate was used to evolve the prescriptors.

The prescriptors were evolved and tested with the same training and testing sets as the global neural net. The prescriptors were fixed fully connected neural networks with two layers of weights. Their weights were initially random, and modified by crossover and mutation. They received the current land-use percentages as their input, and their outputs specified the suggested changed land-use percentages; they were then given to the predictor to estimate the change in ELUC. The outputs were compared to the inputs to calculate the change percentage.

Figure 6.12 demonstrates the progress of evolution towards increasingly better prescriptors, i.e. those that represent better implementations of each tradeoff of the ELUC and change objectives. They represent a wide variety of tradeoffs, and a clear set of dominant solutions that constitute the final Pareto front (red dots). That set is returned to the decision-maker, who can then select the most preferred one to be implemented. Importantly, the evolved Pareto front dominates two linear baselines: one where land is converted to forest from all other types evenly, and another where other land types are converted to forest in a decreasing order of emissions. A closer look revealed that evolution discovered an unexpected strategy: Instead of trying to improve everywhere, as the heuristics did, it identified a smaller number of locations where land-use change had the largest effect, and allocated maximum change to those locations. In other words, it found that it is important to pick your battles! This result suggests that the approach is able to learn and utilize non-obvious opportunities in the domain, and therefore results in better solutions for land use than the obvious heuristics.

6.2.4 Case study: Optimizing NPIs for COVID-19

One example of discovering intelligent decision strategies through neuroevolution is the system for optimizing non-pharmaceutical interventions in the COVID-19 pandemic (Miikkulainen, Francon, et al. 2021). Throughout the pandemic in 2019-2023, governments and decision makers around the world were trying to contain the health and economic impacts of the pandemic by imposing a variety of regulations on the society. Economically the most severe restrictions included school and workplace closings, stay-at-home requirements, and restrictions on public events, gatherings, and domestic and international travel; less severe ones included public information campaigns, testing arrangements, contact tracing, and masking requirements. The approaches were very different around the world, partly

because especially early on it was not clear how effective they each were individually and in combination.

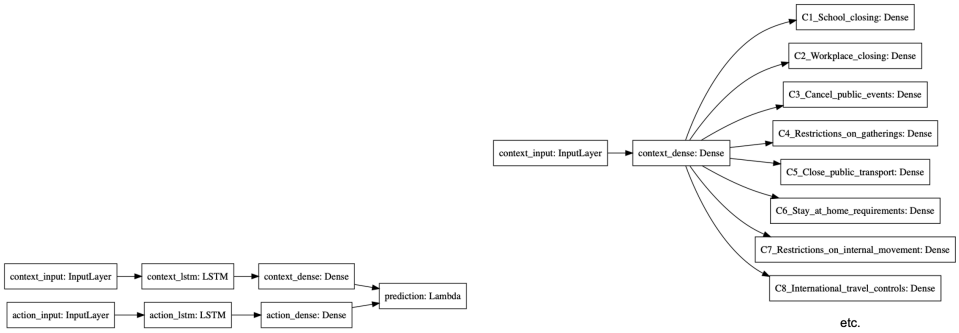
COVID-19 was the first global pandemic that took place in the information age, and data about it became available in vast amounts and almost immediately. It became a major focus of the scientific community (in late 2020, a new paper was submitted to arXiv/bioarXiv on average every 17 minutes), and many approaches were developed to use the data to understand it and cope with it. Most of the approaches were based on existing technology of epidemiological modeling, developed in the early 1900s during and after the major pandemics at that time (Kermack and McKendrick 1927). The idea is to construct differential equations that describe how different populations become susceptible, exposed, infected, and recover or die (SEIR). The models require estimating several parameters, the most important of which is r, the transmission rate. The effect of NPIs can be taken into account by modifying these parameters. More recently these models have been augmented with agent-based modeling approaches and network models, which can extend their granularity almost to an individual person's level (Mark EJ Newman 2002; Venkatramanan et al. 2018). Properly constructed, the models can be accurate and useful in predicting the course of the pandemic. However, estimating the parameters is difficult, and the models are computationally expensive to run.

Much of the community especially early on focused on prediction, i.e. what will happen. The decision makers could then in principle use these predictions to evaluate alternative NPIs and decide what to do about it. Even such communication between the scientists and decision makers turned out difficult, especially in the political climate at the time, but there were several cases where it was effective and resulted in good outcomes (Fox et al. 2022). An interesting question therefore arises: Could optimal intervention policies be discovered automatically using machine learning?

The approach described in the previous section is well suited to this task. The first step is to build the surrogate, i.e. the predictive model that could then be used to evaluate the policy candidates. It turned out that the usual SEIR approaches could not serve this role very well for three reasons: It was difficult to parameterize them for the hundreds of countries and finer-grain locations; it was difficult to parameterize them to model all possible intervention combinations; and the models took too long to run to evaluate the large number of candidate policies that needed to be tested. However, there were enough data available so that it was possible to develop a data-driven approach to prediction: train a neural network to predict the number of cases (or hospitalizations, or deaths) phenomenologically.

The approach was possible because good sources of data existed to construct it. Time series data were available for cases and other indicators for different locations around the world through centralized sources almost daily Disease Control and Prevention 2023. In addition, a major project at Oxford University evaluated government and news outlet sources in order to formalize the NPI policies in effect at these locales (Hale et al. 2020). The NPIs around the world were unified into a representation with 12-20 categories, each with 1-4 stringency levels.

Such data made it possible to use supervised machine learning techniques to form the predictive surrogate model (Figure 6.13*a*). An LSTM neural network with two channels, one for the number of cases, and the other for the NPIs, was trained to predict the cases the next day. As its input, it received the history of the last 21 days, and the predictions were



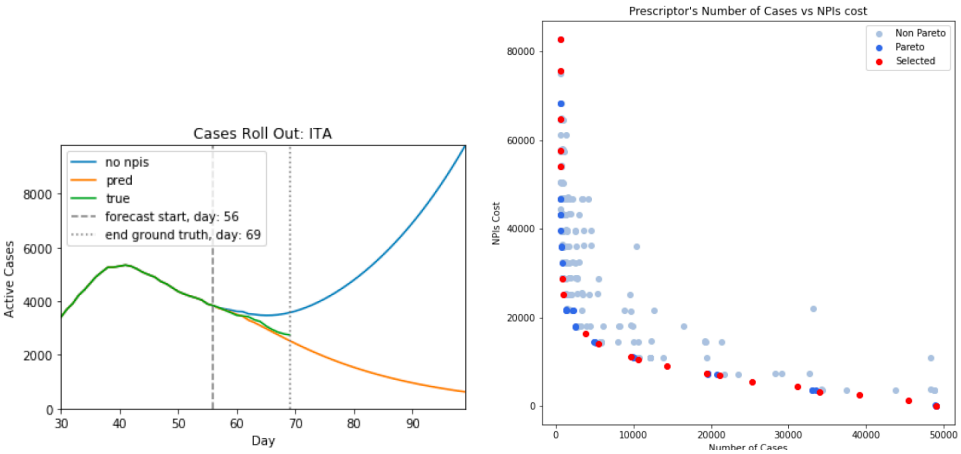
(*a*) Predictor model (*b*) Prescriptor model

Figure 6.13: **Predictive and prescriptive models for discovering nonpharmaceutical interventions (NPIs) in the COVID-19 pandemic.** The predictor is used as a surrogate model for the world in order to evolve prescriptors that implement good NPI strategies. (*a*) The predictor is an LSTM network that receives a 21-day sequence of cases and NPIs as input, and predicts the cases next day. The network is trained with historical data across different countries. During performance, the prediction is looped back to the input, and rolled out indefinitely into the future. (*b*) The prescriptor receives the same sequence of cases and NPIs as input, and prescribes the NPIs for the next day. Since the optimal prescriptions are not known, it is constructed through neuroevolution to reduce both cases and the total stringency of NPIs. Each prescriptor is evaluated through the predictor as the surrogate model. In this manner, the predictor is constructed entirely based on data and is fast enough to evaluate a large number of prescriptor candidates. (Figures from Miikkulainen, Francon, et al. 2021)

looped back into the input so that they could be unrolled indefinitely into the future. The separation made it possible to impose simple constraints on the predictions, such as caps based on the population size of the locale, and that more stringent NPIs should not lead to increases in the number of cases.

The prescriptor models were then evolved to discover good intervention policies (Figure 6.13*b*). Each prescriptor received the same sequence of case numbers and NPIs as its input, and suggested NPIs as its output. These suggestions were input to the predictor which then estimated the number of cases. The cases and NPIs were looped back into the input of both models, and in this manner, the prescriptor was evaluated 90 days into the future. Its performance was measured based on the number of cases as well as the total stringency of the NPIs it suggested. The problem is thus multiobjective, and NSGA-II was used to construct a Pareto front of solutions. Therefore, the end result is a collection of prescriptors on the Pareto front. The idea is that the decision maker can then choose a suitable tradeoff between cases and stringency, i.e. health and economic outcomes.

Note that this problem is a good example of a decision-making task where a surrogate is necessary, for three reasons. First, even if the decision makers could incorporate science into their process, only one decision policy could be implemented at any one time—yet a very large number of alternatives need to be evaluated in the search process. Second, the NPI policies need to be evaluated over a long time during which the world does not stay constant.



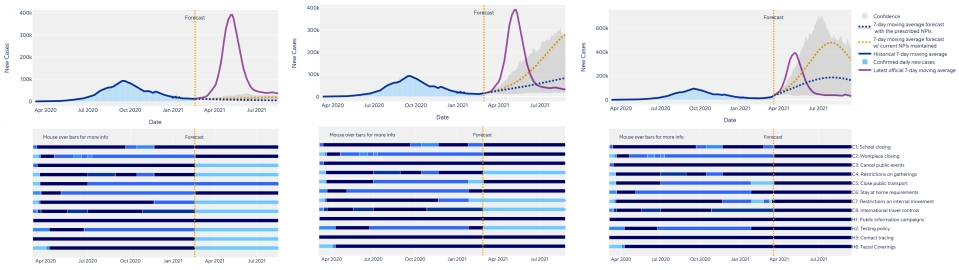
(*a*) Predictor accuracy (*b*) Prescriptor Pareto front

Figure 6.14: **Learned predictors and prescriptors.** (*a*) Given the diverse training data across time and countries, the predictor learned to estimate the number of cases accurately. This example is Italy in July 2020. Given the actual sequence of NPIs as input, it predicted the cases accurately for the next 14 days for which there was data. It also suggested that these NPIs, if maintained, would bring the cases down, but if lifted, and an explosion of cases would result. (*b*) The performance of the final population of prescriptors along the case and cost objectives. The Pareto front evolved strongly towards the bottom left, and in the end offered a set of tradeoffs from which the decision makers can choose. For an animation of the Pareto front, see https://neuroevolutionbook.com/neuroevolution-demos. (Figures from Miikkulainen, Francon, et al. 2021)

The NPIs change over time, the number of cases changes as a result of the NPIs, and also changes differently depending on the stage of the pandemic. The evaluations thus need to be done against a surrogate that is accurate enough to track such changes. Third, simply predicting the most likely outcome is not sufficient; it must also be possible to estimate the uncertainty on the predictions. With a surrogate model, it is possible to estimate the uncertainty in the initial predictions; the evaluation can then be unrolled multiple times to observe the variation in the long term, resulting in confidence bounds.

Throughout the pandemic, from May 2020 through December 2022, the predictor and prescriptor models were trained daily, forming a constantly adapting set of predictions and policies for all locations. The data-driven approach worked surprisingly well in constructing reliable predictors. Different countries implemented different restrictions, and they encountered different phases of the pandemic at different times. Thus, the data was diverse enough so that the predictor learned to evaluate the different policy candidates accurately. These results were confirmed by evaluating the predictions against actual data in various countries at various stages of the pandemic early on. As long as there were no major changes in the NPIs or the pandemic, the predictions tracked the cases well (Figure 6.14*a*).

Similarly, prescriptor evolution discovered a range of effective policies for different stages of the pandemic and for different locations (Figure 6.14*b*). Evaluations with the surrogate model suggest that in many cases they would have resulted in a lower number



(*a*) 2/19/2020 (*b*) 3/1/2020 (*c*) 3/21/2020

Figure 6.15: **The predicted delta surge in India and a presription to avoid it.** (*a*) On 2/19/2020, the cases were decreasing (top plot) and the prescriptors suggested that many NPIs could be lifted (bottom plot, lighter colors). (*b*) The cases were similarly low on 3/1/2020, but there had been delta surges elsewhere and the models predicted a major surge in India if the current NPIs were continued—which was hard to believe at the time. The prescriptors suggested tightening some of them, which could have still avoided a major surge. (*c*) However, more stringent NPIs were only established several weeks later, and by that time even a full lockdown could not have avoided the major surge. In this manner, the models can be used to detect problems early enough when it is still possible to fix them. For an interactive demo, see https://neuroevolutionbook.com/neuroevolution-demos.

of cases and lower economic impact than the actual policies implemented. An interesting pattern of discoveries emerged in this process: The models often discovered principles a few weeks ahead of the time they became widely known. The first such result appeared in May 2020: the models consistently suggested the most stringent restrictions on schools and workplaces. Indeed a few weeks later results came out suggesting that the virus was transmitted most effectively in such closed spaces where people stayed in contact for several hours every day. In September 2020 the suggestions changed, focusing on gatherings and travel restrictions, but suggesting less stringent restrictions for schools. Indeed measures had been taken at schools wrt. separation, ventilation, dividers, and masks that made it possible to keep them open in a more safe manner.

Perhaps the most significant demonstration of the power of the approach took place in March 2021, during the delta variant surge. The models predicted a huge explosion of cases in India, which was surprising because India had had the pandemic under control until then and there was no indication that anything was wrong. However, the models had seen delta surges elsewhere, and apparently recognized that the NPIs at the time made it vulnerable. Even though it was difficult to believe the models, they were correct. If the recommendations had been followed, much of the surge could have been avoided (Figure 6.15).

On the other hand, the models were much less successful in coping with the omicron surge. It was indeed different in that it happened very rapidly all over the world—there was not enough time for the models to get to see it in some countries, and then apply it to others. It also turned out that in 2022 it no longer made sense to train the models from all the available data. Different NPIs were used: there was better testing, tracing, and masking, and fewer restrictions on work, school, and travel. Also, people behaved differently in 2022

compared to 2020. In many locations, they did not adhere to the restrictions the same way, and also masking, testing, and vaccinations made it less necessary to do so. Therefore, it was better to train the models with less but more recent data. On the other hand, this result again emphasized that it is important to train the predictor together with the prescriptor; in that manner, they can both adapt to the changing world.

The NPI optimization application, as described above, was primarily a technology demo, but it has already had a significant impact. In a couple of cases it was also used to inform actual policy decisions, such as the school openings in Iceland in the Fall of 2021. A major effort in mainstreaming the approach was the XPRIZE Pandemic Response Challenge in December 2020-March 2021 (XPRIZE 2023; Cognizant AI Labs 2023). Over 100 teams around the world participated in creating predictors and prescriptors for the pandemic. The general setup and the data sources were the same, but the approaches varied widely. The winning teams were successful not only in terms of performance, but also in communicating the results with decision makers. Most recently, Project Resilience (ITU 2023), a project led by the ITU agency of the United Nations, is an attempt to build on these successes further and extend to other challenges such as the climate change. In this manner, over time, it is possible that the surrogate optimization approach in general, and neuroevolution in particular, will gradually become widely used in coping with a variety of problems in decision making in society.

An interactive demo of the NPI optimization system is at... It allows going back in time and evaluating the model's suggestions, comparing them to actuals, and modifying them to see the effects. Also the code prepared for the XPRIZE competition is available at... Using that starting point, it is possible to develop further models for the pandemic dataset and others.

6.2.5 Leveraging human expertise

Recent applications of supervised learning have demonstrated the power of learning the statistics of large numbers of labeled examples, and various reinforcement learning and evolutionary optimization approaches have reached super-human performance in many game-playing domains without much human involvement. However, there are many domains where humans have significant expertise. Incorporating such expertise in learning could provide a better starting point, allowing it to find better solutions in complex tasks, and also solutions that may be easier and safer to deploy.

Neuroevolution provides a natural way to incorporate such knowledge into creative problem-solving. Human solutions can be encoded in equivalent neural networks to form the initial population, which is then evolved further to take advantage both of the knowledge and machine discovery.

A method called RHEA (Realizing Human Expertise through AI) was developed for this purpose (Meyerson et al. 2024). It consists of four phases: (1) Define the problem in a manner such that diverse expertise can be applied to it. (2) Gather the solutions from the experts. (3) Distill the solutions into a population of equivalent neural networks. (4) Evolve the neural network population to discover improved solutions.

Let us illustrate the approach first in a synthetic domain illustrated in Figure 6.16. The problem is defined as one where a subset of policy interventions $a_1, a_2 ... a_n$ an need to be selected for different contexts $c_1, c_2 ... c_m$ to optimize utility ϕ and cost ψ. Assume there are

three expert solutions are available: two specialists for c_1 and c_2 and a generalist that can be applied across all contexts. They can be distilled into a common grid representation where black in cell (c_i, a_j) indicates choosing an action a_j for context c_i. This population of three solutions can then be evolved to obtain better solutions.

Let the utility be defined as

$$\phi(c,A) = \begin{cases} 1, & \text{if } c = c_1 \wedge A = \{a_1, a_2\} \\ 2, & \text{if } c = c_1 \wedge A = \{a_1, a_2, a_3, a_4, a_5\} \\ 3, & \text{if } c = c_1 \wedge A = \{a_1, a_2, a_3, a_4, a_5, a_6\} \\ 4, & \text{if } c = c_2 \wedge A = \{a_1, a_2, a_3, a_4, a_5, a_6\} \\ 5, & \text{if } c = c_2 \wedge A = \{a_1, a_2, a_3, a_4, a_6\} \\ 1, & \text{if } c = c_2 \wedge A = \{a_3, a_4, a_5\} \\ 1, & \text{if } A = \{a_7, a_8, a_9, a_{10}\} \\ 0, & \text{otherwise.} \end{cases} \tag{6.29}$$

and the cost ψ be the number of actions in the solution. The Pareto front resulting from RHEA is illustrated on top of Figure 6.16. Some of the solutions are found by recombining existing expert solutions, e.g. by adding a_3, a_4, a_5 to a_1, a_2 in c_1. Importantly, evolution can also innovate beyond the experts, e.g. by adding a_6 to this solution. It can also refine solutions by removing actions that are redundant or detrimental, such as a_5 in c_2, and by incorporating knowledge from the generalist solution, i.e. $a_7..a_{10}$ for $c_3...c_7$.

Interestingly, other methods cannot take advantage of such mechanisms. For instance Mixture-of-Experts (MoE; Masoudnia and Ebrahimpour 2014) can utilize different experts for different contexts (as shown at the bottom of Figure 6.16), but cannot form recombinations of them, or innovations or refinements. Its Pareto front therefore falls far short of that of evolution. Similarly, Weighted Ensemble solutions (Dietterich et al. 2002) can only choose a single combination of experts that is then applied to all contexts, which results in even less effective Pareto front.

Note also that it would be difficult for evolution alone to find a good Pareto front, i.e. starting from random solutions instead of the experts. There is little information in partial solutions that allows constructing them gradually, and evolution would thus be looking for needles in a haystack. Indeed, experimentally RHEA discovers the entire optimal Pareto front reliably whereas evolution does not, especially when the number of actions increases.

This synthetic example thus illustrates how evolution can take advantage of expert knowledge, how it can improve solutions beyond such knowledge, and how these abilities are unique to evolution as compared to standard machine learning approaches. Do these insights carry over to large real-world domains?

To demonstrate the real-world power of RHEA, it was implemented in the XPRIZE Pandemic Response domain mentioned in the previous section. In Phase 2 of the competition, a total of 169 different prescriptors were submitted. They were constructed with different methods such as epidemiological modeling, decision rules, statistical methods, gradient-based optimization, and evolution; some of them also utilized auxiliary data sources, and some focused on specific locations. This set of prescriptors was thus quite diverse, representing diverse human expertise. Several studies in psychology, social science, and business

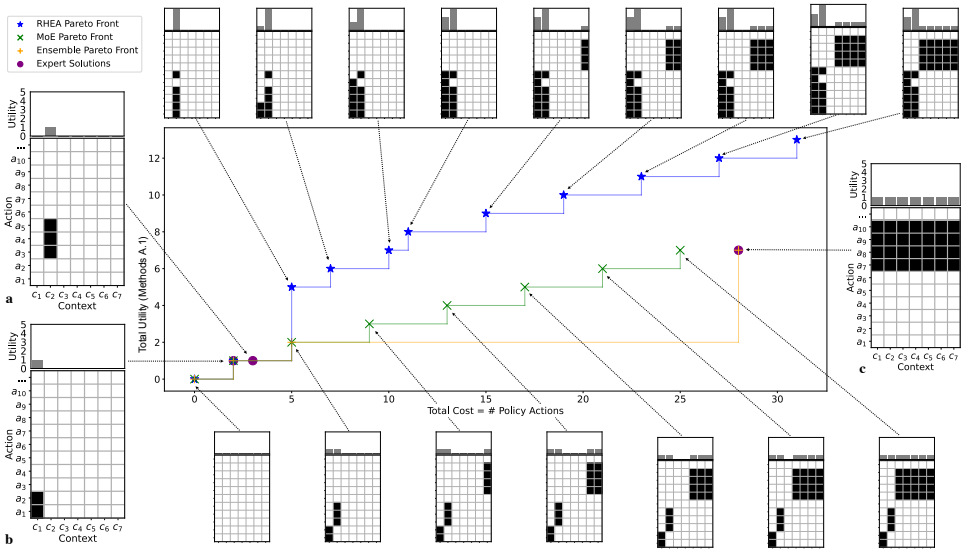
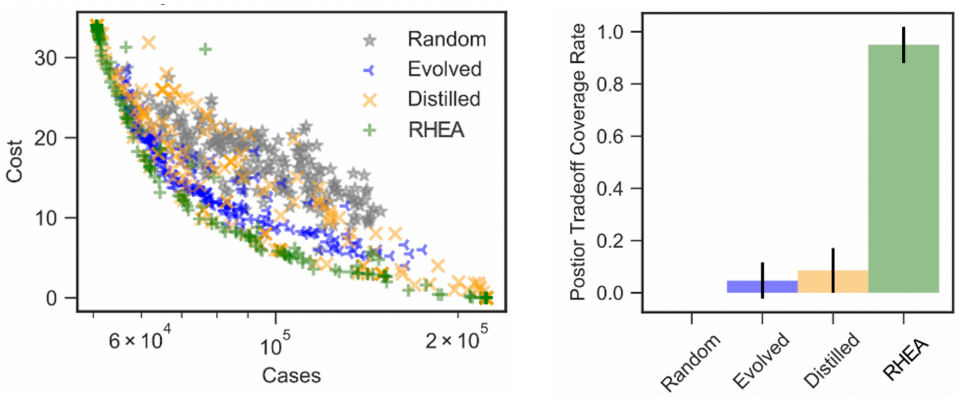


Figure 6.16: **RHEA leveraging expert solutions through evolution, compared to Mixture-of-Experts (MoE) and Weighted Ensemble.** Several solutions may include different good ideas; the challenge is to form a combined solution that takes advantage of all of them. In this synthetic example, the plots in the middle show the Pareto fronts for each method: RHEA in blue ⋆, MoE in green ×, and Weighted Ensemble in yellow +; in addition, the original expert solutions are shown in purple •. The structure of each solution is visualized as a grid that identifies which actions (row) are used in each context (columns. On the left are the two original specialist solutions **a** and **b**, and on the right, the original generalist solution **c**. The solutions on the RHEA Pareto front are on top, and those for MoA in the bottom. Whereas MoE and Weighted Ensemble can utilize the knowledge in the expert solutions only in a limited way, RHEA can recombine, add innovations, and remove redundancies and detrimental elements to construct superior solutions. Whereas such solutions would be difficult to evolve from a random initial population, RHEA thus harnesses the latent potential in expert solutions, and finds the optimal Pareto front reliably. (Figures from Meyerson et al. 2024)

suggest that diversity in human teams leads to improved decision-making (Rock and Grant 2016). The question is: Can we use AI (i.e. neuroevolution) to take advantage of this diversity of human expertise?

The XPRIZE competition provided a convenient framework for the first two phases. The distillation was done by training an autoregressive neural network with gradient descent to mimic the behavior of each solution created by human experts. Training examples were created by querying the prescriptor with a comprehensive sampling of the Oxford data set. Evolution was done through the same ESP approach as described in the previous section. That is, the latest predictor at the time was used as the surrogate, and neural networks optimized the case and cost objectives as before.

Remarkably, the results exceeded all expectations (Figure 6.17). The RHEA Pareto front pushed significantly further down and to the left than the Pareto front consisting of the best



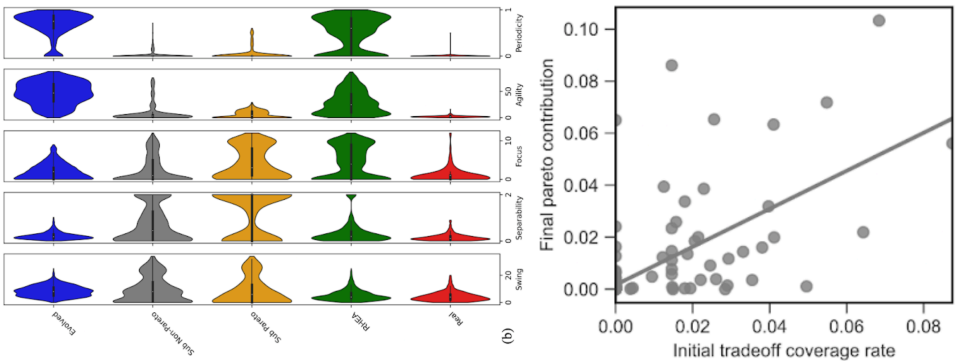
(*a*) Pareto fronts (*b*) Human-preferred solutions

Figure 6.17: **Combining human expertise and machine discovery in NPI optimization.** The recombination and mutation operators in evolution are well-suited for combining, refining, and extending existing ideas. (*a*) The RHEA Pareto front dominates both the solutions created by human experts (Distilled), as well as solutions evolved from a random initial population. (*b*) Given the human decision makers' preference for mid-range tradeoffs, RHEA's solutions would be selected nearly always. These results demonstrate that neuroevolution can be used to take advantage of human expertise, resulting in solutions that are better than both those of humans and evolution alone. (Figures from Meyerson et al. 2024)

solutions created by human experts, as well as the Pareto front resulting from the evolution from initially random neural networks. In other words, RHEA evolution was more powerful than either human expertise or evolution from scratch alone. Moreover, the RHEA solutions dominated especially in the areas of the front that mattered: Given the human decision-makers' preference for mid-range tradeoffs, they would be likely to select RHEA's solutions over those of other methods nearly 100% of the time.

It is interesting to evaluate what RHEA actually discovered differently from humans and machines alone. Figure 6.18(*a*) characterizes the policies along five dimensions: The range of their stringency (swing), whether they utilize different phases (separability), number of IPs used (focus), how often the IPs change (agility), and whether they utilize weekly changes (periodicity). The policies are characterized for RHEA, evolution-only, and submitted solutions, as well as the actual policies implemented in the world during the pandemic.

Several interesting observations can be made from this comparison. First, in terms of swing and separability, the submitted solutions had more variability than policies in the real world, suggesting that human experts were exploring opportunities to improve. However, RHEA's solutions were more similar to the real world, although RHEA also discovered that extreme separability could sometimes be useful. In this manner, RHEA did discover that the human expert's innovations were not always productive. Second, in terms of focus, RHEA's solutions were more similar to the submitted solutions, and quite different from the real-world solutions. In this manner, it utilized the expert solutions' tendency to focus on a small number of NPIs. Third, in terms of agility and periodicity, RHEA differed from both



(*a*) Dimensions of NPI strategies (*b*) Performance vs. contribution

Figure 6.18: **Characterizing the discovered NPI policies.** The policies can be characterized in five dimensions, revealing similarities and differences between approaches. (*a*) RHEA's policies were similar to the submitted ones in terms of focus, but differed in four other dimensions. In terms of swing and separability, it found solutions similar to those implemented in the real world, but in terms of agility and periodicity, a potential new opportunity that both human experts and real-world decision-makers missed. In this manner, RHEA can leverage both human expertise and machine creativity. (*b*) Performance (in terms of hypervolume) of the submitted solutions vs. their contributions to the final Pareto front. While better solutions generally contribute more, there are many solutions that do not perform well but end up contributing a lot (those in the upper left area). This result highlights the value of soliciting diverse expertise even if some of it is not immediately useful: Methods such as RHEA can then be used to realize their latent potential. (Figures from Meyerson et al. 2024)

submitted and real-world solutions, utilizing more frequent variations as well as weekly periodicity. The solutions that were evolved from a random starting point were similar along these two dimensions, suggesting that they were indeed discovered through machine creativity. Such solutions tend to be more difficult to implement in the real world, although in some cases they had been (e.g. for a time in Portugal and France). In this sense, RHEA discovered a potential opportunity that both real-world decision-makers and human experts' solutions had missed. The conclusion is that RHEA can indeed utilize ideas from solutions created by human experts as well as develop its own in order to construct the best possible policies.

It is also interesting to characterize how RHEA discovered the best solutions, by analyzing their evolutionary history. Some such solutions can be traced back to only a single beneficial crossover of two submitted ancestors, while others were constructed in a more complex process involving several ancestors. Usually the crossovers were systematic, i.e. resulted in offspring whose case-stringency tradeoff was in-between the two parents. It is also interesting to measure the contribution of each ancestor to the solutions in the final Pareto front, i.e. how much of their genetic encoding was found in those best solutions (Figure 6.18*b*). As expected, submitted ancestors that performed well generally contributed

more, but there are also many ancestors that made outsize contributions through the evolutionary process. This observation demonstrates why it is so useful to solicit diversity of expertise, even when some of it is not immediately useful. Neuroevolution methods such as RHEA can then be used to realize their latent potential.

The NPI optimization example demonstrates the power of RHEA in combining human expertise and machine creativity through neuroevolution. The approach can be applied to many other domains as well where such diverse expertise is available. It can be further combined with techniques for trustworthiness, such as interactive exploration and confidence estimation. Neuroevolution can thus play a crucial role in taking advantage of intelligent decision-making in the real world.

Note that in RHEA, human expertise is treated as a black box. This approach makes it possible to utilize such expertise in any form, distilled into a common neural network representation. However, sometimes expertise is available explicitly in the form of rules, examples, and advice. Such knowledge can be incorporated into neuroevolution by modifying the evolved networks directly, as will be discussed in Section 8.2. It is a different way of utilizing human expertise in neuroevolution.

Interestingly, distillation can also be useful in the other direction, i.e. by taking a neural network that performs well as a black box, and then evolving a set of rules to replicate its performance (Shahrzad, Hodjat, and Miikkulainen 2024). Rule sets are transparent and interpretable, and in this manner, it may be possible to explain how the network performs. In particular with RHEA, this approach may make it possible to characterize the initial expert solutions in a uniform manner, and further identify what new knowledge evolution discovers to improve them. Neuroevolution can thus work synergetically with rule-set evolution to make both human and AI designs explainable.

6.3 Chapter Review Questions

1. **Levels of Behavior:** Describe the different levels of behavior that neuroevolution aims to optimize, from low-level control to high-level decision strategies. Provide an example of a success story for each level.
2. **Robust Behavior:** What are some challenges in evolving robust behaviors in dynamic or unpredictable environments? Discuss methods like trajectory noise, coevolution, or symmetry evolution that address these challenges.
3. **Simulation to Reality Transfer:** Explain how neuroevolution can be adapted to bridge the ”reality gap” between simulations and the physical world. What role does noise, stochasticity, and modern robotics simulators play in this process?
4. **Behavioral Switching:** Why is switching between high-level strategies more challenging than low-level control adjustments in neuroevolution? Provide examples of fractured decision boundaries and interleaved/blended behaviors that illustrate these challenges.
5. **Fractured Strategies and Network Design:** Explain how specific network design choices, such as using radial basis functions or cascaded refinement, can address the challenge of discovering fractured decision boundaries in domains like half-field soccer.

6. **Multimodal Task Division:** Discuss the role of preference neurons in discovering and implementing multimodal behaviors. How does this approach enable neuroevolution to discover surprising and effective strategies, such as in the Ms. Pac-Man example?
7. **Surrogate Modeling:** What is the role of surrogate models in discovering decision strategies with neuroevolution? Discuss how they enable exploration and evaluation in domains where real-world experimentation is infeasible.
8. **Evolutionary Surrogate-Assisted Prescription (ESP):** Describe the ESP process for decision-making. How does co-learning between predictors and prescriptors contribute to automatic regularization and curricular learning?
9. **COVID-19 NPI Optimization:** In the context of optimizing non-pharmaceutical interventions during the COVID-19 pandemic, how did the ESP approach combine predictive and prescriptive modeling to discover effective policies? What were the advantages of this data-driven method over traditional epidemiological models?
10. **Human Expertise in RHEA:** Explain how RHEA (Realizing Human Expertise through AI) incorporates human expertise into neuroevolution. How does it utilize diverse expert solutions to discover superior decision strategies, and what unique advantages does it provide over other methods like Mixture-of-Experts?