# Technical Appendix
## Pilot Experimentation for Safe Exploration and Side-Effects Avoidance

**C Henrik Åslund, Alessio Lomuscio**

Imperial College London
c.aslund19@imperial.ac.uk, a.lomuscio@imperial.ac.uk

## A Examples

**Agricultural Science.** Cells can correspond to different regions of the world separated in different ways. That a pesticide experiment in America is unlikely to affect Eurasia is an example of separation by distance. An experiment in one body of water is unlikely to affect another body of water. Likewise, an experiment on one island is unlikely to affect another island. There are also artificial separations. A mesocosm, for example, is a laboratory sealed off from the world. In particular, it is a miniature replication of a natural ecosystem. Testing chemical pesticides is one kind of experiment, but introducing new species and seeing if they can protect crops is another kind of experiment. In that case, potential adverse side-effects would be that the new species becomes invasive and damages the local ecosystem. As for regulators, we can let them model regulatory agencies, e.g. EPA and EFSA, where the cell would be the region of the US or the EU, respectively.

**Recommender Engines.** Cells correspond to different kinds of users. Cellular independence models that a recommendation given to one user does not affect another user. Different kinds of users have been argued to be more susceptible to misinformation such as children or unemployed adults can be included in $\mathbf{G}$. Moderators are the regulators in this case and could also be included in $\mathbf{G}$. The reason is that we model the moderator as interacting with the system as a user. That some regulators may always be silent models that not all users are moderators.

**Medical Regulations.** Regulations from *Fundamentals of Clinical Trials* (Friedman et al. 2015) include the following.

1. 'A small number of participants, usually three, are entered sequentially at a particular dose. If no [...] toxicity is observed, the next predefined dose is used.'
2. 'If unacceptable toxicity is observed, in any of the three participants, additional participants, usually three, are treated at the same dose.'
3. 'If no further toxicity is seen, the dose is escalated to the next higher dose. If additional unacceptable toxicity is observed, then the dose escalation is terminated.'
4. 'Once a drug or biologic has undergone some minimal investigation, it should be available to those with life-threatening conditions, should they desire it.'
5. 'Pregnant women are often excluded from drug trials [...] Similarly, investigators would probably exclude [...] people with a recent history of gastric bleeding.'



## B Additional Figures

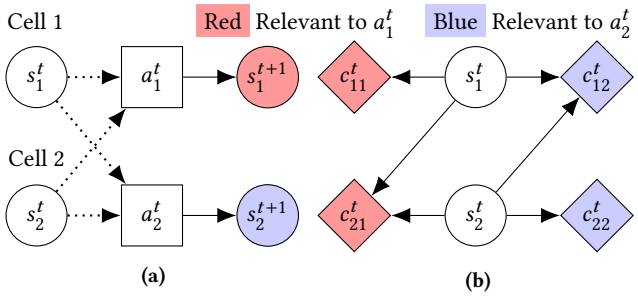


Figure B.1: The core properties of cellular MDPs in the style of causal influence diagrams (Everitt et al. 2021). There are $n = 2$ cells with $i = 1$ above and $i = 2$ below. The colour red illustrates what is relevant to $a_1^t$ and blue what is relevant to $a_2^t$. Panel (a))illustrates that each intracellular action affects each transition independently, but the policy may depend on information from both cells. Panel (b) illustrates that the regulator in cell 1 depends on $s_1^t$; their view on safety regarding cell 2 depends on $s_2^t$ as well.

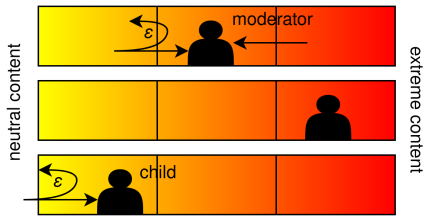


Figure B.2: Polarisation in recommender engines—the reset variant. Arrows are intracellular actions pushing users in one direction or the other, and $\varepsilon$ is a small probability of the push having the opposite effect.

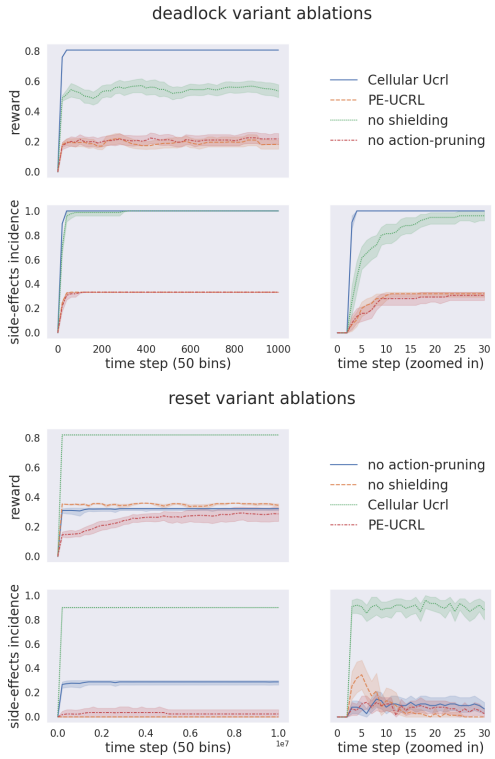


Figure B.3: Reward and side-effects incidence over time. Averages and standard deviations for $\geq$ 20 samples are shown. Algorithm names abbreviate algorithm–regulations pairs.

# C Algorithms

**UCRL.** Theoretically efficient algorithms for average RL include upper confidence RL algorithms such as UCRL2 (Jaksch, Ortner, and Auer 2010) and KL-UCRL (Filippi, Cappé, and Garivier 2010). Upper confidence RL algorithms explore by assuming the most optimistic values of the confidence intervals over transition probabilities and rewards. They differ in how these confidence intervals are computed. However, they share that the policies are computed using extended value iteration. Let $\mathbf{P}$ be the set of transition functions, and let $\tilde{R}$ be the maximal reward function within the confidence set. Extended value iteration is given by the recurrent function

$$V\{h{+}1\}(\mathbf{s}) := \max_{\substack{\mathbf{a}\in\mathbf{A}(\mathbf{s})\\ \tilde{P}\in\mathbf{P}}} \sum_{\mathbf{s}'\in\mathbf{S}} \left[\tilde{R}(\mathbf{s},\mathbf{a},\mathbf{s}') + V\{h\}(\mathbf{s}')\right]\tilde{P}(\mathbf{s}'|\mathbf{s},\mathbf{a})$$

$$V\{h\}(\mathbf{s}) := 0 \qquad \text{(C.1)}$$

where $\{h\}$ denotes the $h$th iteration. The time complexity of each iteration is in $\mathrm{O}(|\mathbf{S}|^2|\mathbf{A}|)$. The iterations stop when $\max_{\mathbf{s}\in\mathcal{S}} \Delta\{h\}(\mathbf{s}) - \min_{\mathbf{s}\in\mathcal{S}} \Delta\{h\}(\mathbf{s}) \leq 1/t[k]$, where $\Delta\{h\}(\mathbf{s}) := V\{h+1\}(\mathbf{s}) - V\{h\}(\mathbf{s})$ and $t[k]$ is the time step that episode $k$ started on.

**CELLULAR UCRL.** To take into account transfer, CELLULAR UCRL uses a couple of new counters. Instead of applying to states and actions, these counters apply to intracellular state–action pairs. Let $v_\#$ be a counter such that $v_\#[k](s_\#, a_\#)$ is the number times $(s_\#, a_\#) \in \mathbf{S}_\# \times \mathbf{A}_\#$ has been observed across both cells and time since the start of the $k$th on-policy phase. Let $N_\#[k] := v_\#[k-1] + N_\#[k-1]$. $v_\#$ is updated at each time step, and when *switchPhase*, defined below, CELLULAR UCRL proceeds into the off-policy phase.

$$\mathit{swichPhase} \equiv (v_\rightarrow[k](\mathbf{s},\mathbf{a}) \geq \max\{1, N_\rightarrow[k](\mathbf{s},\mathbf{a})\})$$

$$\text{where } \begin{cases} v_\rightarrow[k](\mathbf{s},\mathbf{a}) := \min\limits_{(s_\#,a_\#)\in\{(s_i,a_i)\}_{i\in[n]}} v_\#[k](s_\#,a_\#) \\ N_\rightarrow[k](\mathbf{s},\mathbf{a}) := \min\limits_{(s_\#,a_\#)\in\{(s_i,a_i)\}_{i\in[n]}} N_\#[k](s_\#,a_\#) \end{cases}$$

In the off-policy phase, we modify the calculation of the confidence set $\mathbf{CM}$. When CELLULAR UCRL is based on UCRL2 (with confidence level $\delta \in ]0,1[$), the possible transition functions $\tilde{P}$ in $\mathbf{CM}$ are defined by

$$\|\tilde{P}(\cdot|\mathbf{s},\mathbf{a}) - \hat{P}[k](\cdot|\mathbf{s},\mathbf{a})\|_1 \leq \sqrt{\frac{14|\mathbf{S}|\log(2|\mathbf{A}|t[k]/\delta)}{\max\{1, N_\rightarrow[k](\mathbf{s},\mathbf{a})\}}}$$

$$\text{where } \hat{P}[k](\mathbf{s}'|\mathbf{s},\mathbf{a}) = \textstyle\prod_{i\in[n]} \hat{P}_\#[k](s_i'|s_i,a_i).$$

CELLULAR UCRL then updates the policy through extended value iteration, Equation (C.1), and proceeds into the next on-policy phase.

Algorithm C.1: ACTION-PRUNING

**Input**: last state $\mathbf{s}$, last action $\mathbf{a}$, current state $\mathbf{s}'$, side-effects $\mathbf{c}$, confidence set $\mathbf{CM}$, partial prior knowledge $pk$
**Output**: pruned et $\mathbf{CM}^-$, Boolean *updatePolicy*

$\mathbf{CM}^- \leftarrow \mathbf{CM}$ and $\mathbf{A}^-(\cdot) \leftarrow \mathbf{A}(\cdot)$ **if** first call to algorithm
*updatePolicy* $\leftarrow$ *false*
**for** $i, j \in [pk.n]$ **do** ▷ aggregate
  remove $\tilde{\mathbf{C}}$ s.t. $\mathit{safe}\in\tilde{\mathbf{C}}(\mathbf{s})$ from $\mathbf{CM}^-$ **if** $c_{ij}{=}\mathit{unsafe}$
  remove $\tilde{\mathbf{C}}$ s.t. $\mathit{unsafe}\in\tilde{\mathbf{C}}(\mathbf{s})$ from $\mathbf{CM}^-$ **if** $c_{ij}{=}\mathit{safe}$
**for** $j \in [pk.n]$ **do** ▷ remove (basic case)
  **if** $\tilde{\mathbf{C}}$ exists s.t. $\mathit{unsafe} \in \tilde{\mathbf{C}}(\mathbf{s})$ in $\mathbf{CM}^-$ **then**
    *updatePolicy* $\leftarrow$ *true*
    remove $\tilde{P}$ s.t. $\tilde{P}(\mathbf{s}'|\mathbf{s},\mathbf{a}) > 0$ from $\mathbf{CM}^-$
    remove $a_j$ from $\mathbf{A}^-_\#(s_j)$
**for** $j \in [pk.n]$ **do** ▷ remove (corner case)
  $path_j \leftarrow \emptyset$ **if** first call to algorithm
  add $(s_j,a_j,s_j')$ to $path_j$ **if** $|\mathbf{A}^-_j(\mathbf{s}')|{\leq}1$ **else** $path_j{\leftarrow}\emptyset$
  **if** $\tilde{\mathbf{C}}$exists s.t.$\mathit{safe}\in\tilde{\mathbf{C}}(\mathbf{s})$in$\mathbf{CM}^-$or$|\mathbf{A}^-_\#(s_j')|{=}0$ **then**
    *updatePolicy* $\leftarrow$ *true*
    **for** $(s_\#, a_\#, s_\#') \in path_j$ **do**
      remove $\tilde{P}$ s.t. $\tilde{P}(\mathbf{s}'|\mathbf{s},\mathbf{a}) > 0$ from $\mathbf{CM}^-$
      remove $a_j$ from $\mathbf{A}^-_\#(s_j)$ (i.e. $\mathbf{A}^-_\#(s_j) \leftarrow \emptyset$)

**PE-UCRL (no shield).** Algorithm C.1 shows the subroutine ACTION-PRUNING in detail. First, it takes the side effects $\mathbf{c} \sim \mathbf{C}(\mathbf{s})$ as input and aggregates them. The aggregation subroutine assumes consistent regulators from Definition 2 in the main paper. The idea is that, if there exists a cell that is safe according to any regulator in any cell, then

that is the same as that cell being safe according to all regulators. This part has time complexity $\mathrm{O}(n^2)$.

The main idea behind action-pruning is to, from the action space, prune an action taken in a specific state immediately after the state–action pair has been observed to be unsafe. There are corner cases caused by the last action being removed from a state. For example, consider the cellular MDP with the transition structure

$$\underrightarrow{a_i' \leftarrow s_i \rightarrow a_i \rightarrow s_i' \rightarrow a_i \rightarrow s_i'' \rightarrow a_i}$$

in cell $i$. Intracellular states $s_i, s_i'$ are safe and intracellular state $s_i''$ is unsafe. Hence, intracellular action $a_i$ ought to be pruned from $\mathbf{A}_{\#}(\mathbf{s}_i)$ and $\mathbf{A}_{\#}(\mathbf{s}_i')$; only intracellular action $a_i'$ should be kept. The solution is given by the second half, corner cases, of *ActionPruning*. In the basic case, the time complexity is $\mathrm{O}(n)$, but it increases to $\mathrm{O}(n \max_{i\in[n]} |\mathbf{S}_i|)$ when also accounting for corner cases.

The output $\mathbf{CM}^-$ is then used as input to planning in the form of extended value iteration, see Equation (C.1). If the output $updatePolicy \equiv true$ but $switchPhase \equiv false$, PE-UCRL (with or without shield) proceeds to the off-policy phase without incrementing $k$.

**PE-UCRL (no pruning).** This algorithm is identical to PE-UCRL except that ACTION-PRUNING is not called. It is used for ablation studies.

**NATION-LIKE.** NATION-LIKE models that different nation-states (or other kinds of jurisdictions) have different regulations that constrain different kinds of exploratory policies. If adverse side effects are observed in one jurisdiction, other jurisdictions can make use of this knowledge without being affected themselves, which provides some level of safety. NATION-LIKE is easiest motivated through an example. Consider the Thalidomide disaster (Johnson, Stokes, and Arndt 2018). Different nations have different regulations regarding the approval of new drugs. As a simplification, consider different nations to have different probabilities of approving a new drug. What happened in the Thalidomide disaster was that some countries accepted the drug while others did not. Adverse side effects were observed in the nations that did and not in the others. However, the other nations still learned the lesson that Thalidomide was unsafe. (Similar stories apply to e.g. pesticides and additives to plastics.) As for recommender engines, the question may be more about which recommender engines are allowed in what countries. China, e.g., tends to have much more stringent regulations. This causes many foreign recommender engines to be unable to operate within that nation. Other examples where jurisdictions learn from one another include vaccination policies, GMO policies, and data protection policies. Note that this is an example of an uncoordinated (decentralised) approach to exploration that has some safety features. Therefore, nation-like algorithms can work as a kind of baseline. It represents the level of safety we can expect to achieve if we do not implement any specific safety features.

For simplicity, the implementation is based on PE-UCRL (no shield). The regulatory constraints take two arguments: One is the probability of a new intracellular policy being admitted (this could also be specified as a list with different values for different cells). In the experiments, we used the propbability $0.2$. The other is the number of time steps between each off-policy phase (this could, e.g., correspond to a five-year plan). Finally, the policies are maximised greedily (there is no coordination on how to explore). In the experiments, these updates happen every $50$ time step.

**AUP.** The contribution in AUP (Turner, Ratzlaff, and Tadepalli 2020) is a certain regulariser that can be used to shape the reward function of any reinforcement learning algorithm. The idea of the regulariser is to randomly generate a number of auxiliary reward functions. The Q-functions ($Q_{auxR_l}$) under the policies that maximise these auxiliary reward functions are calculated. The shaped rewards satisfy

$$R_{shaped}(\mathbf{s},\mathbf{a}) = R(\mathbf{s},\mathbf{a}) - \lambda \frac{\sum_l Q_{auxR_l}(s,a) - V^{\boldsymbol{\pi}^{\text{init}}}_{auxR_l}(\mathbf{s})}{\sum_l V^{\boldsymbol{\pi}^{\text{init}}}_{auxR_l}(\mathbf{s})},$$

where $\lambda$ is a hyperparameter. The reason why this might avoid side effects is that the algorithm is incentivised to *keep its options open*. Turner et al. implement this loss function in Q-learning and PPO algorithms, whereas we apply it to CELLULAR UCRL instead. The regulatory constraints contain information about $\lambda$ as well as the number of auxiliary reward functions. In the experiments, we use $\lambda = 1$ and $l \in [10]$, i.e. 10 auxiliary reward functions.

**ALWAYSSAFE/PSO.** There are two safe exploration algorithms making use of factored MDPs: AlwaysSafe (Simão, Jansen, and Spaan 2021) and PSO (Farquhar, Carey, and Everitt 2022). Adapted to the cellular MDP model, these two algorithms become indistinguishable.

AlwaysSafe assumes prior knowledge of the transitions for a subset of the variables in the MDP. This subset contains all variables that are safety-relevant. In the cellular MDP model, we do not assume full knowledge of the transitions for any subset of variables. We merely assume knowledge that if variables (or combinations thereof) correspond to cells, then (1) their transitions are independent and (2) transitions in cells are identical. AlwaysSafe cannot make use of (2) and making use of (1) makes it much weaker—we explain below. To adapt AlwaysSafe to cellular, MDPs, we assume worst-case probabilities for the transitions. This means that if there is a cell that could enter an unsafe state (regardless of the policy), then any update to the corresponding intracellular policy would be unsafe. To conclude, in cellular MDPs, AlwaysSafe can only regard cells as either admissible for experiments or not admissible for experiments. Therefore, all the modifications go through the regulatory constraints. The regulatory constraints take the form of lists of classes of cells that could potentially turn unsafe. None of the cells that could potentially turn unsafe have their corresponding intracellular policies updated. All other intracellular policies can be updated freely.

PSO assumes that the regulatory constraints designate some variables as *delicate*. The algorithm is designed not to cause changes in the state transitions in these delicate cells. Adapted to the cellular MDP model, this means that certain

cells, let us call them *delicate cells* should not be subjected to other intracellular policies (exploratory) than the initial safe policy. In conclusion, these delicate cells correspond perfectly to the cells that could potentially turn unsafe from the cellular AlwaysSafe algorithm. Therefore, the cellular PSO algorithm is indistinguishable.

# D Proofs

**(Restated) Theorem 1** (Safety)**.** *With probability* $\geq 1-\delta$, PE-UCRL *satisfies* $\Phi$ *from Definition 5. Its time complexity between every time step in the off-policy phase is polynomial in* $\mathrm{O}(n|\mathbf{S}|^2|\mathbf{A}|)$. *Its time complexity between every time step in the on-policy phase is in* $\mathrm{O}(n^2 + \max_{i\in[n]}|\mathbf{S}_i|)$.

*Proof sketch.* First, we prove satisfaction of the regulatory constraints. The verification is sound if the true $CM$ is in $\mathbf{CM}$ (Puggelli et al. 2013). In brief, we apply Theorem 2.1 from Weissman et al. (2003) and the union bound to get that

$$\mathbb{P}\left[\begin{array}{c}\|P(\cdot\mid\mathbf{s},\mathbf{a})-\hat{P}(\cdot\mid\mathbf{s},\mathbf{a})\|_1\\ \geq\\ \sqrt{\frac{14|\mathbf{S}|\log(2n|\mathbf{A}|t/\delta)}{\max\{1,N_{\rightarrow}[k](\mathbf{s},\mathbf{a})\}}}\end{array}\right]^{(n-1)(t[k]-1)} \leq \sum_{\nu=1}\frac{\delta}{20t^7n|\mathbf{S}||\mathbf{A}|}.$$

Noting that the righthand side is $<\delta/(20t^6|\mathbf{S}||\mathbf{A}|)$ and that

$$\sqrt{\frac{14|\mathbf{S}|\log(2n|\mathbf{A}|t/\delta)}{\max\{1,N_{\rightarrow}[k](\mathbf{s},\mathbf{a})\}}} \geq \sqrt{\frac{14|\mathbf{S}|\log(2|\mathbf{A}|t/\delta)}{\max\{1,N_{\rightarrow}[k](\mathbf{s},\mathbf{a})\}}},$$

we can apply the same reasoning as Jaksch et al. (2010) did for UCRL2. Therefore, the true $CM$ is in $\mathbf{CM}$ with probability $\geq 1-\delta$. This means that PE-SHIELD outputs a safe policy if the verification would return *true*. The assumption on the initial policy being safe means it always would for some number of iterations. Second, the complexities are determined by combining the results from Appendix C. □

**(Restated) Corollary 1.1.** *Assume: The cellular MDP is state transience invariant. Then, the time complexity of* PE-UCRL *between every time step in the on-policy phase is in* $\mathrm{O}(n^2)$.

*Proof sketch..* If $CM$ is state transience invariant, then the corner cases part of ACTION-PRUNING is never used. □

**Definition D.1** (Expected Adverse Side-Effects)**.** Let $pk{+}C(\mathbf{S}_{\#}^{\subseteq})$ be an object equal to $pk$ except that side-effects are reported for all $\mathbf{s}\in\mathbf{S}_{\#}^{\subseteq}$. Adverse side-effects are expected for $\mathbf{S}_{\#}^{\subseteq}$ if there is some on-policy phase $k$ such that the RL agent acts as if $prior = pk{+}\mathbf{C}(\mathbf{S}_{\#}^{\subseteq})$.

**Definition D.2** (Overhead)**.** Let $\mathbf{v}_{\rightarrow}[k]\in\mathbb{N}^{|\mathbf{S}|}$ be defined such that its sth element is the visitation count of the pair $(\mathbf{s},\boldsymbol{\pi}^{\Phi}(\mathbf{s}))$ in the $k$th on-policy phase (i.e. $\mathbf{v}_{\rightarrow}[k] \coloneqq [v_{\rightarrow}(\mathbf{s},\boldsymbol{\pi}^{\Phi}(\mathbf{s}))]_{\mathbf{s}\in\mathbf{S}}$). Let

$$\mathbf{P}[k] \coloneqq [[P(\mathbf{s}'\mid\mathbf{s},\pi^{+}(\mathbf{s}))]_{\mathbf{s}\in\mathbf{S}}^{\mathsf{T}}]_{\mathbf{s}'\in\mathbf{S}} \text{ and}$$
$$\mathbf{P}^{\Phi}[k] \coloneqq [[P(\mathbf{s}'\mid\mathbf{s},\pi^{\Phi}(\mathbf{s}))]_{\mathbf{s}\in\mathbf{S}}^{\mathsf{T}}]_{\mathbf{s}'\in\mathbf{S}}.$$

Let $\mathbf{w}[k]\in\mathbb{R}^{|\mathbf{S}|}$ be a constant vector such that $\|\mathbf{w}\|_{\infty}\leq D/2$. The overhead is defined as

$$oh_{pk+\mathbf{C}(\mathbf{S}_{\#}^{\subseteq})}[k] \coloneqq \mathbf{v}_{\rightarrow}[k](\mathbf{P}[k]-\mathbf{P}^{\Phi}[k])\mathbf{w}[k].$$

Let $K\in\mathbb{N}$ be an upper bound on the number of on-policy phases before time $T$. We define $oh_{prior}(T) \coloneqq \sum_{k\in[K]} oh_{prior}[k]$.

**Lemma 1** (Regret)**.** *Assume: The cellular MDP is state transient invariant and communicating. Then, with probability* $\geq 1-\delta$,

$$\underset{(\text{PE-UCRL},pk+\mathbf{C}(\mathbf{S}_{\#}))}{\mathrm{Regret}(T)} \leq \begin{array}{c}34D|\mathbf{S}|\sqrt{T|\mathbf{A}|\log(T/\delta)}\\ +\\ oh_{pk+\mathbf{C}(\mathbf{S}_{\#})}(T)\end{array}$$

*Proof sketch..* In line with Jaksch et al. (2010), we divide the proof into three steps: (1) We bound the regret when the true $CM$ is not in the confidence set $\mathbf{CM}$. (2) We bound the regret when it is. (3) We combine these bounds. Step (1) follows the same line of reasoning as the proof of Theorem 1. Step (3) is identical to the proof by Jaksch et al. Similarities to that proof can be leveraged because of the following: That cellular MDP is state transience invariant and communicating ensures that the action-pruned cellular MDP $CM^{-}$ is also communicating. That all adverse side effects are expected, i.e. that $pk{+}\mathbf{C}$ applies to all of $\mathbf{S}_{\#}$, ensures that the exploration, which is only dependent on $P$ and $R$, is sufficient.

As for Step (2), first note that $N_{\rightarrow}[k](\mathbf{s},\mathbf{a})\geq N[k](\mathbf{s},\mathbf{a})$ for all $\mathbf{s}\in\mathbf{S}$ and $\mathbf{a}\in\mathbf{A}$. Consequently, we can make the substitution

$$\sqrt{\frac{14|\mathbf{S}|\log(2|\mathbf{A}|t[k]/\delta)}{\max\{1,N_{\rightarrow}[k](\mathbf{s},\mathbf{a})\}}} \leq \sqrt{\frac{14|\mathbf{S}|\log(2|\mathbf{A}|t[k]/\delta)}{\max\{1,N[k](\mathbf{s},\mathbf{a})\}}}$$

and reuse most of the proof by Jaksch et al. The exception is regarding the introduction of the true transition matrix into the upper bound. We add $\mathbf{0}=-\mathbf{P}[k]+\mathbf{P}[k]-\mathbf{P}^{\Phi}[k]+\mathbf{P}^{\Phi}[k]$ resulting in

$$\tilde{\mathbf{v}}[k](\mathbf{P}^{+}[k]-\mathbf{I})\mathbf{w}[k] = \underbrace{\tilde{\mathbf{v}}[k](\mathbf{P}[k]-\mathbf{P}^{\Phi}[k])\mathbf{w}[k]}_{\text{old terms}} + \overbrace{\tilde{\mathbf{v}}[k](\mathbf{P}^{+}[k]-\mathbf{P}[k])\mathbf{w}[k]+\tilde{\mathbf{v}}[k](\mathbf{P}^{\Phi}[k]-\mathbf{I})\mathbf{w}[k]}^{\text{new term}}$$

where $\mathbf{I}\in\mathbb{R}^{|\mathcal{S}|\times|\mathbf{S}|}$ is an identity matrix, and $\mathbf{P}^{+}[k]$ is the most optimistic choice of transition matrix given the set of transition functions $\tilde{P}[k]$ under the target policy $\boldsymbol{\pi}^{+}[k]$. We bound the old terms identically to Jaksch et al. The new term is equal to $oh^{\Phi}[k]$. Summing this term over $k$ lets us replace it with $oh^{\Phi}(T)$. □

**Definition D.3** (Regular Visitation)**.** A state $\mathbf{s}$ is regularly visited if

$$\textstyle\sum_{\mathbf{a}\in\mathbf{A}(\mathbf{s})} N_{\rightarrow}[k](\mathbf{s},\mathbf{a})\rightarrow\infty \text{ when } k\rightarrow\infty,$$

An space $\mathbf{S}^{\subseteq}\subseteq\mathbf{S}$ is regularly visited if all states in $\mathbf{S}^{\subseteq}$ are regularly visited (likewise for $\mathbf{S}_{\#}^{\subseteq}\subseteq\mathbf{S}_{\#}$).

**Lemma 2** (Sublinear Overhead)**.** *Assume: The cellular MDP is state transient invariant and communicating. The prior knowlegde* $pk + \mathbf{C}(\mathbf{S}_{\#}^{\subseteq})$ *strikes a suitable trade-off between explorability and expressivity.* $\mathbf{S}_{\#}^{\subseteq}$ *is regularly visited. Then,* $oh_{pk+\mathbf{C}(\mathbf{S}_{\#}^{\subseteq})}(T)$ *approaches a constant as* $T \to \infty$*.*

*Proof sketch.*. An equivalent formulation of the lemma is that there exists a $k$ such that, for all $k' > k$ and all $\mathbf{s} \in \mathcal{S}$ one of the following conditions holds: (1) $v_{\to}[k'](\mathbf{s}, \boldsymbol{\pi}^{\Phi}[k'](\mathbf{s})) = 0$, or (2) for all $\mathbf{s} \in \mathcal{S}$, $P(\mathbf{s}' \mid \mathbf{s}, \boldsymbol{\pi}^{+}[k'](\mathbf{s})) - P(\mathbf{s}' \mid \boldsymbol{\pi}^{\Phi}[k'](\mathbf{s})) = 0$.

By assumption, adverse side effects are expected for some $k'$ due to the sequence $\boldsymbol{\pi}^{\Phi}[1], ..., \boldsymbol{\pi}^{\Phi}[k'], ...$ We claim that it is also the case that adverse side effects would be expected if a sequence $\boldsymbol{\pi}^{\Phi}[1], ..., \boldsymbol{\pi}^{\Phi}[k'], \boldsymbol{\pi}^{+}[k'+1], ...$ were in use. The core idea is that the set of regularly visited states for the latter is a subset of those for the former. From the assumption that $\Phi$ is reasonable and the inner workings of PE-SHIELD, ift follows that, for all $k' > k$, $\pi_i^{+}[k'] = \pi_i^{\Phi}[k']$ for some $i$. Now, we argue that as the set of such cells $i$ expands, the intracellular policies do not change their visitations of intracellular states. This follows from the assumption that the reward function is monotonically increasing and the inner workings of extended value iteration. (A potential complication would be *jumbling* in PE-SHIELD. Note, however, that the probability of jumbling quickly approaches 0.)

Our next claim is that, for all sufficiently late $k'' > k'$, $\boldsymbol{\pi}^{+}[k'']$ would be safe if it were used. The core idea is to fix $k'$ in a certain way. First, let $k'$ be fixed such that the regularly visited states for the policies after $k'$ is a subset of those visited by the policy at $k$. Second, let $k'$ earliest be the point after which adverse side effects are expected. *ActionPruning* removes the unsafe actions from extended value iteration. Thereby, all subsequent target policies would be safe. It can be shown that this fixation is possible.

Finally, *PE-Shield* gradually replaces $\pi_i^{\Phi}$ with $\pi_i^{+}$ for an expanding set of cells $i$. So, for all states that are regularly visited, condition (2) holds. For the remaining states, condition (1) holds. □

**(Restated) Theorem 2** (Convergence)**.** *Assume: The cellular MDP is state transient invariant and communicating. The prior knowledge* $pk$ *strikes a suitable trade-off between explorability and expressivity. Reporting is complete. Then, with probability* $\geq 1 - \delta$*, for large* $T \in \mathbb{N}$*,*

$$\underset{(\text{PE-UCRL}, pk)}{\text{Regret}}(T) \in \tilde{\text{O}}(D|\mathbf{S}|\sqrt{T|\mathbf{A}|})$$

*Proof sketch.* We combine lemmas 1 and 2 and apply the assumption from Definition 10. That reporting is complete and the cellular MDP is communicating imply that there exists a $k$th on-policy phase such that (PE-UCRL, $pk$) and (PE-UCRL, $pk+\mathbf{C}(\mathbf{S}_{\#})$) are equivalent for all $k' > k$. The eventual equivalence is guaranteed by Equation (7) in the main paper. □

# E Environments

The reset and deadlock variants have several things in common. For both, it holds that: There are 3 cells. There are 3 intracellular states (i.e. 27 states). There are 3 intracellular actions (i.e. 27 actions). A graphical representation of the deadlock variant is shown in Figure B.2. In the recommender engine example, the intracellular states correspond to a spectrum from neural to extreme (e.g. political polarization). Each intracellular action can either turn the intracellular state 1 degree more extreme or 1 degree more neutral or remain at the current degree (e.g. news recommendations that influence a user one way or another). The initial state is the leftmost and the initial policy for the RL agent is to remain there. The higher rewards exist around the right end. The total reward is nonlinear in the contributions from each cell. (In particular, it is the 2-logarithm of the sum of the rewards from each cell.) The reward function is known to the RL agent from initialisation. There is one regulator (e.g. a moderator), which is very likely to report side effects as along as they are not in the rightmost intracellular state. There is one cell label (it corresponds to a child using the recommender engine).

**Reset variant.** In the reset variant (1) any state can be reached from any other (assumption 1), and (2) any policy from any state will eventually bring the system back to the initial state. Hence the name reset. The motivation is to show certain convergence properties. Figure B.2 would represent the reset variant were it not for the arrow representing radicalising recommendations. Because of (1), it is in principle possible for an RL agent to reach a *safe optimum*, i.e., the maximum reward subject to the constraint that no unsafe states are visited. This can be contrasted with the *unconstrained optimum*, i.e., the maximum reward without any regard to either safety or regulatory constraints. Because side-effects and rewards are positively correlated in these environments, algorithms without regard for safety will accumulate more rewards.

**Deadlock variant.** In the deadlock variant, a user can become *radicalised*, see the middle user in Figure B.2. That is, if the user ends up in a state (representing their political beliefs) that is extreme enough, they can never return to a more neutral position. In other words, both (1) and (2) from the reset variant are violated. The violation of (1) means that it may not be possible for an agent to find the safe optimum in the worst-case scenario. In the worst case, the agent gets stuck. Therefore, the goal in this environment is to merely keep an acceptable level of safety violations throughout the run. If the resulting policy gets more reward than the initial policy, that is good, but guaranteed convergence is not possible even in principle.

# F Additional Results

**Ablations.** The results of an ablation study is shown in Figure B.3.

(PE-UCRL (no shielding), n/a) performs well in the reset variant. However, it does maximise side effects in the deadlock variant.

(PE-UCRL (no action-pruning), $[\bigcirc \bigcirc (№_{all} \geq 1)]_{\leq 0.5}$) satisfies the specification of the regulatory constraints at all times. However, side-effects incidence increases and plateaus over time.

(CELLULAR UCRL, n/a) is an ablation of both the shield and the action pruning. It combines the worst of both worlds.

**Machine.** All experiments were run on a Linux Machine. The Linux kernel release was 6.0.12-100.fc35.x86_64. The kernel version was #1 SMP PREEMPT_DYNAMIC Thu Dec 8 16:53:55 UTC 2022. It had 56 cores. It had 504 GB memory and 16 GB swap.

## References


Everitt, T.; Carey, R.; Langlois, E. D.; Ortega, P. A.; and Legg, S. 2021. Agent Incentives: A Causal Perspective. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 11487–11495. AAAI Press.

Farquhar, S.; Carey, R.; and Everitt, T. 2022. Path-Specific Objectives for Safer Agent Incentives. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 9529–9538. AAAI Press.

Filippi, S.; Cappé, O.; and Garivier, A. 2010. Optimism in reinforcement learning and Kullback-Leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 115–122. IEEE.

Friedman, L. M.; Furberg, C. D.; DeMets, D. L.; Reboussin, D. M.; and Granger, C. B. 2015. *Fundamentals of clinical trials*. Springer.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *J. Mach. Learn. Res.*, 11: 1563–1600.

Johnson, M.; Stokes, R. G.; and Arndt, T. 2018. *The Thalidomide Catastrophe: How it Happened, Who was Responsible and why the Search for Justice Continues after More than Six Decades*. Onwards & Upwards.

Puggelli, A.; Li, W.; Sangiovanni-Vincentelli, A. L.; and Seshia, S. A. 2013. Polynomial-Time Verification of PCTL Properties of MDPs with Convex Uncertainties. In Sharygina, N.; and Veith, H., eds., *Computer Aided Verification - 25th International Conference, CAV 2013, Saint Petersburg, Russia, July 13-19, 2013. Proceedings*, volume 8044 of *Lecture Notes in Computer Science*, 527–542. Springer.

Simão, T. D.; Jansen, N.; and Spaan, M. T. J. 2021. AlwaysSafe: Reinforcement Learning without Safety Constraint Violations during Training. In Dignum, F.; Lomuscio, A.; Endriss, U.; and Nowé, A., eds., *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, 1226–1235. ACM.

Turner, A. M.; Ratzlaff, N.; and Tadepalli, P. 2020. Avoiding Side Effects in Complex Environments. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Weissman, T.; Ordentlich, E.; Seroussi, G.; Verdu, S.; and Weinberger, M. J. 2003. Inequalities for the L1 deviation of the empirical distribution.