Appendix A: Proof of Theorem 4.1

Theorem 4.1. *The counterfactual ETT is empirically estimable for arbitrary action-choice dimension (i.e., $|X| = k$ for $k \geq 2$) when agents condition on their intent $I = i$ and estimate the response Y to their final action choice $X = a$.*

Proof. We start by writing the corresponding ETT expansion and note that,

$$E[Y_{X=a}|X = i] \tag{1}$$
$$= \sum_{i'} E[Y_{X=a}|X = i, I = i']P(I = i'|X = i) \tag{2}$$
$$= \sum_{i'} E[Y_{X=a}|I = i']P(I = i'|X = i) \tag{3}$$
$$= \sum_{i'} E[Y_{X=a}|I_{x=a} = i']P(I = i'|X = i) \tag{4}$$
$$= \sum_{i'} E[Y|do(X = a), I = i']P(I = i'|X = i) \tag{5}$$
$$= \sum_{i'} E[Y|do(X = a), I = i']1(i' = i) \tag{6}$$
$$= E[Y|do(X = a), I = i] \tag{7}$$

Eq. (2) expands the ETT using the law of total probability to sum over all intent conditions. Eq. (3) follows from the conditional independence $Y_x \perp\!\!\!\perp X|I$ that holds, allowing us to remove $X = i$. Eq. (4) follows because $I_x = I$ given that $(I \perp\!\!\!\perp X)G_x$, where G_x is the interventional submodel where all causal parents of X are severed (as represented by the counterfactual antecedent notation). Eq. (5) is a notational re-arranging because all variables (Y_x and I_x) are in terms of the interventional submodel M_x (and thus G_x), licensing us to express the quantity using the $do(x)$ notation. Eqs. (6,7) follow from the fact that, observationally, an agent's final arm choice will always coincide with their intent (i.e., $P(i|x) = 1\ \forall i = x, 0$ otherwise), which nullifies all summed expressions where the two differ. □

Appendix B: Additional Simulations

To illustrate that the success of TS^{RDC+} is not limited to the simulations of the main text, we performed experiments across a wide swath of different payout parameterizations. In each simulation below, TS^{RDC+} experiences statistically significantly less cumulative regret than its competitors, demonstrating its capacity to successfully navigate arbitrary parameter configurations.

Though it would be intractable to exhaustively test performance in all payout parameterizations, we have attempted to follow a principled approach in which optimal arms are arranged in permutations of their relation to intent. The following is an overview of our parameter choices:

1. **Greedy Casino:** the optimal arm choice is *never* the same as intent (see paper; not in appendix).
2. **Paradoxical Switching:** the optimal arm choice is *always* the same as intent, but paradoxically, the arm with the highest experimental payout $E[Y_{x_1}] = 0.5$ is only the optimal choice in a single intent condition.
3. **Generous Casino:** the optimal arm choice is *always* the same as intent.
4. **Odd Man Out:** a single intent arm is suboptimal, the others are optimal.
5. **Sometimes Switch:** the optimal arm choice is the same as intent in two intent-conditions, but different in the other two.

Note: for all parameterizations that follow, $P(D = 1) = P(D = 0) = 0.5$, $P(B = 1) = P(B = 0) = 0.5$, and the agent's intent is decided by the structural equation: $X \leftarrow f_X(B, D) = B + 2 * D$.

Paradoxical Switching

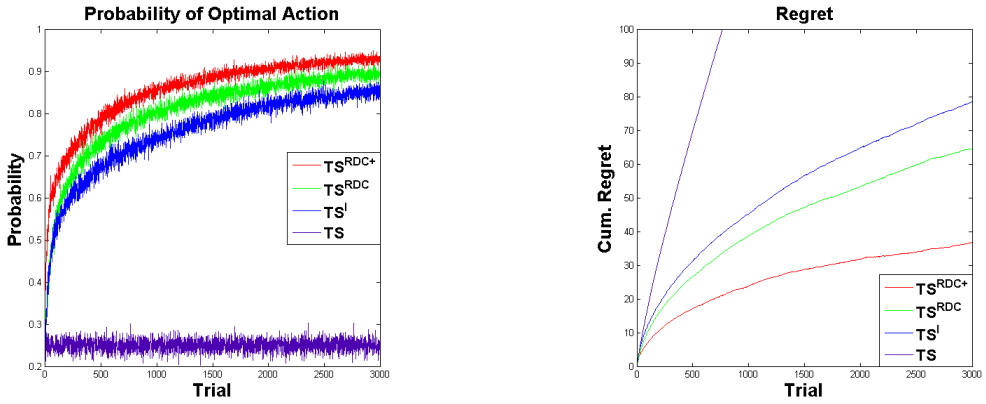

Figure 1: Plots of TS variant performances in the Paradoxical Switching scenario.

(a)	$D = 0$		$D = 1$	
$E[y_1\|X, B, D]$	$B = 0$	$B = 1$	$B = 0$	$B = 1$
$X = 0$	*0.90	0.20	0.45	0.45
$X = 1$	0.30	*0.40	0.50	0.40
$X = 2$	0.10	0.35	*0.60	0.35
$X = 3$	0.10	0.10	0.30	*0.60

(b)	$E[y_1\|X]$	$E[y_1\|do(X)]$
$X = 0$	0.90	0.50
$X = 1$	0.40	0.40
$X = 2$	0.60	0.35
$X = 3$	0.60	0.20

Table 1: (a) Payout rates decided by reactive slot machines as a function of arm choice X, sobriety D, and machine conspicuousness B. Players' natural arm choices under D, B are indicated by asterisks. (b) Payout rates according to the observational, $E(y_1|X)$, and experimental $E(y_1|do(X))$, distributions, where $Y = y_1$ represents winning (shown in the table).

The Paradoxical Switching parameterization exemplifies a curious scenario wherein $E[Y_{x_1}] = 0.5 > E[Y_{x'}]\ \forall x' \neq x_1$, but for which x_1 is the optimal arm choice in only one intent condition ($I = x_1$). Agents unempowered by RDC will face a paradox in that the arm with the highest experimental payout is not always optimal. Again, TS^{RDC+} experienced significantly less regret ($M = 36.91$) than its chief competitor, TS^{RDC}, ($M = 64.70$), $t(1998) = 22.43, p < .001$.

The Generous Casino

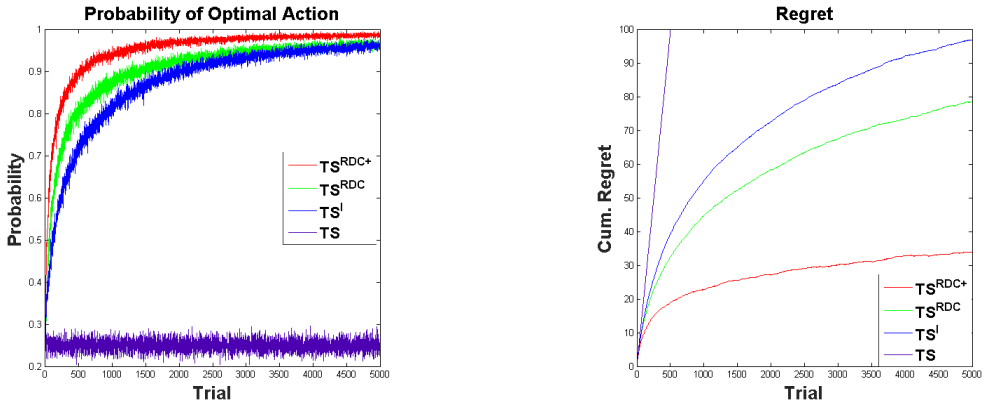

Figure 2: Plots of TS variant performances in the Generous Casino scenario.

(a)	$D=0$		$D=1$	
$E[y_1\|X,B,D]$	$B=0$	$B=1$	$B=0$	$B=1$
$X=0$	*0.60	0.20	0.50	0.30
$X=1$	0.30	*0.60	0.20	0.50
$X=2$	0.50	0.30	*0.60	0.20
$X=3$	0.20	0.50	0.30	*0.60

(b)	$E[y_1\|X]$	$E[y_1\|do(X)]$
$X=0$	0.60	0.40
$X=1$	0.60	0.40
$X=2$	0.60	0.40
$X=3$	0.60	0.40

Table 2: (a) Payout rates decided by reactive slot machines as a function of arm choice X, sobriety D, and machine conspicuousness B. Players' natural arm choices under D, B are indicated by asterisks. (b) Payout rates according to the observational, $E(y_1|X)$, and experimental $E(y_1|do(X))$, distributions, where $Y = y_1$ represents winning (shown in the table).

The Generous Casino parameterization is the dual of the Greedy Casino in which the agent's intent is always the optimal choice. This parameterization illustrates that there is no inherent "distrust" of intent in the $RDC+$ algorithm, and furthers its generalizability. In this simulation, TS^{RDC+} experienced significantly less regret $(M = 33.99)$ than its chief competitor, TS^{RDC}, $(M = 78.64)$, $t(1998) = 25.35, p < .001$.

Odd Man Out

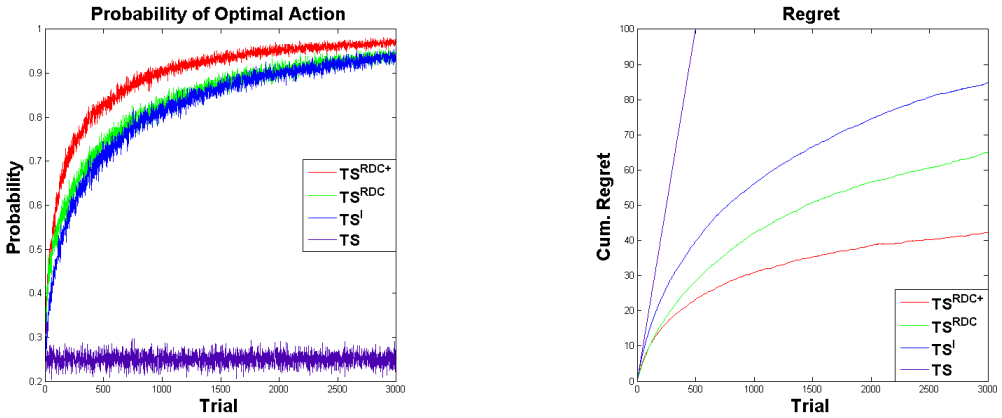

Figure 3: Plots of TS variant performances in the Odd Man Out scenario.

(a)	D = 0		D = 1	
$E[y_1\|X,B,D]$	B = 0	B = 1	B = 0	B = 1
$X=0$	*0.40	0.40	0.50	0.30
$X=1$	0.20	*0.60	0.30	0.50
$X=2$	0.50	0.30	*0.60	0.20
$X=3$	0.20	0.50	0.30	*0.60

(b)	$E[y_1\|X]$	$E[y_1\|do(X)]$
$X=0$	0.40	0.40
$X=1$	0.60	0.40
$X=2$	0.60	0.40
$X=3$	0.60	0.40

Table 3: (a) Payout rates decided by reactive slot machines as a function of arm choice X, sobriety D, and machine conspicuousness B. Players' natural arm choices under D, B are indicated by asterisks. (b) Payout rates according to the observational, $E(y_1|X)$, and experimental $E(y_1|do(X))$, distributions, where $Y = y_1$ represents winning (shown in the table).

The Odd Man Out parameterization explores the scenario where a single intent arm is suboptimal. It illustrates the capacity of $RDC+$ to account for asymmetrical exceptions in parameterizations. In this simulation, TS^{RDC+} experienced significantly less regret $(M = 52.39)$ than its chief competitor, TS^{RDC}, $(M = 78.72)$, $t(1998) = 12.75, p < .001$.

Sometimes Switch

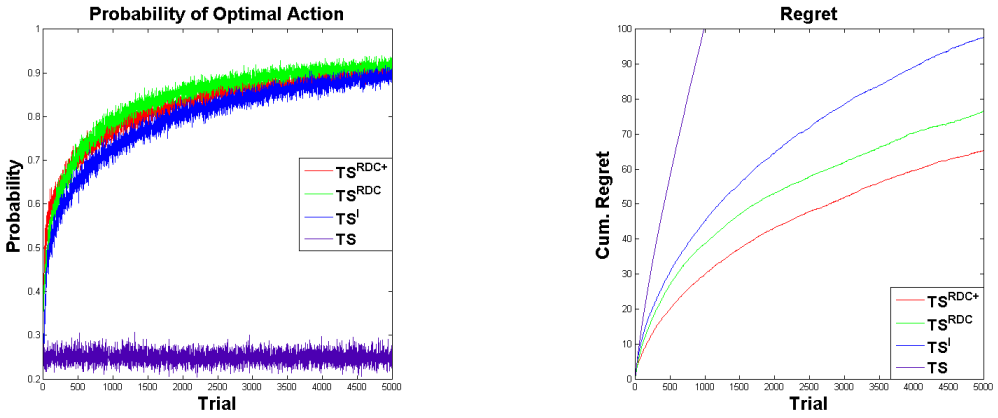

Figure 4: Plots of TS variant performances in the Sometimes Switch scenario.

(a)	$D = 0$		$D = 1$	
$E[y_1 \mid X, B, D]$	$B = 0$	$B = 1$	$B = 0$	$B = 1$
$X = 0$	*0.90	0.20	0.45	0.45
$X = 1$	0.40	*0.30	0.50	0.40
$X = 2$	0.25	0.40	*0.40	0.35
$X = 3$	0.10	0.10	0.10	*0.50

(b)	$E[y_1 \mid X]$	$E[y_1 \mid do(X)]$
$X = 0$	0.90	0.50
$X = 1$	0.30	0.40
$X = 2$	0.40	0.35
$X = 3$	0.50	0.20

Table 4: (a) Payout rates decided by reactive slot machines as a function of arm choice X, sobriety D, and machine conspicuousness B. Players' natural arm choices under D, B are indicated by asterisks. (b) Payout rates according to the observational, $E(y_1|X)$, and experimental $E(y_1|do(X))$, distributions, where $Y = y_1$ represents winning (shown in the table).

The Sometimes Switch scenario hosts two intent arms that are optimal and two that are not. Furthermore, it disrupts any symmetrical relationships between and within the observational and experimental payouts (as they are all different within arms). In this simulation, TS^{RDC+} experienced significantly less regret $(M = 65.41)$ than its chief competitor, TS^{RDC}, $(M = 76.64)$, $t(1998) = 5.48, p < .001$.