Reinforcement Learning for Digital Health Interventions in the Dyadic Setting

HeartSteps

Susan Murphy

Oralytics

ADAPTS HCT

MiWaves

DOI

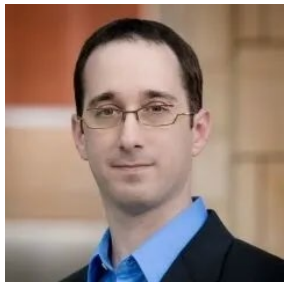
d3c optimizing adaptive interventions, by design

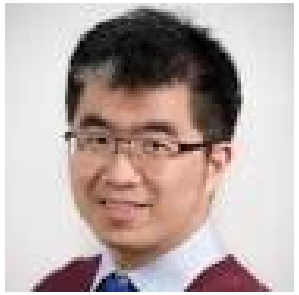
Kempner INSTITUTE For the Study of Natural & Artificial Intelligence at Harvard University

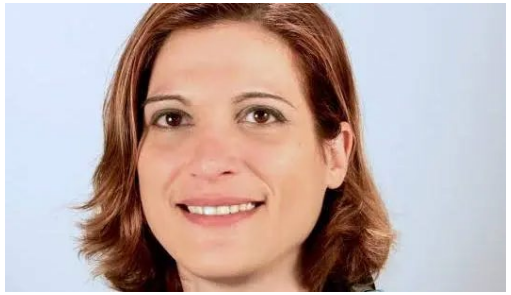
Ziping Xu

ADAPTS HCT

Acknowledgements



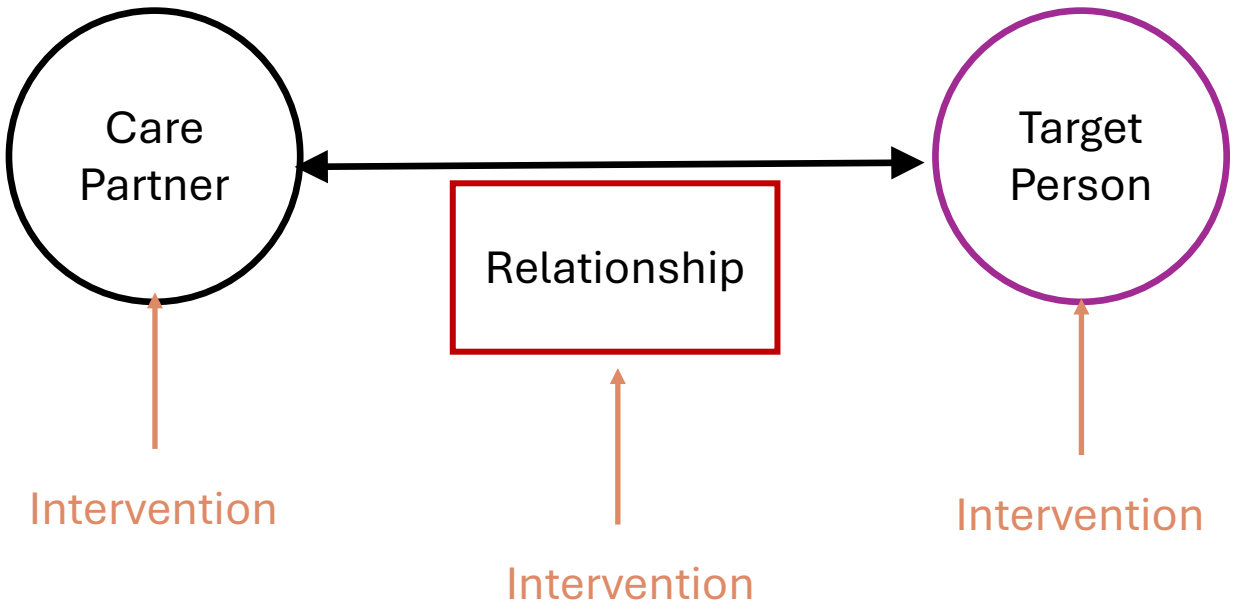




HARVARD UNIVERSITY

A Starting Point: Dyadic Setting

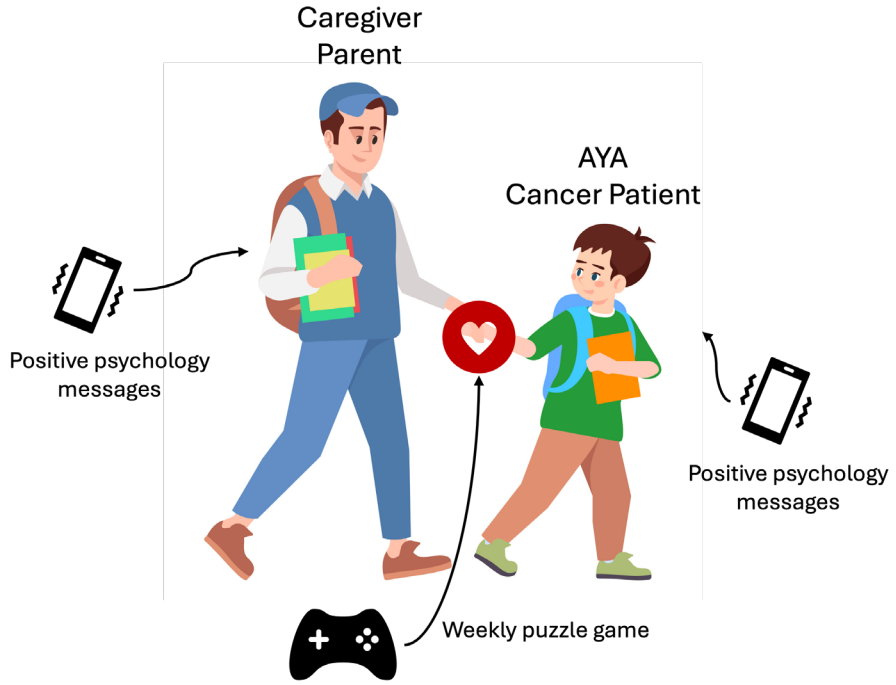
- Social network with two individuals



GOAL: Support Target Person's Health Outcome

ADAPTS-HCT

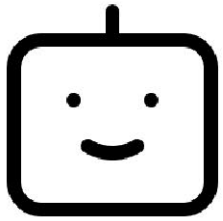
- Dyads
  - **AYA** (adolescent and young adults)
    - Undergone blood and marrow transplant
  - **Care partner** (e.g., a parent)



- Goal: enhance AYA's **medication adherence**

Interventions for Dyadic Structure

Offer?



Not offer?

Agent

AYA

Positive emotions

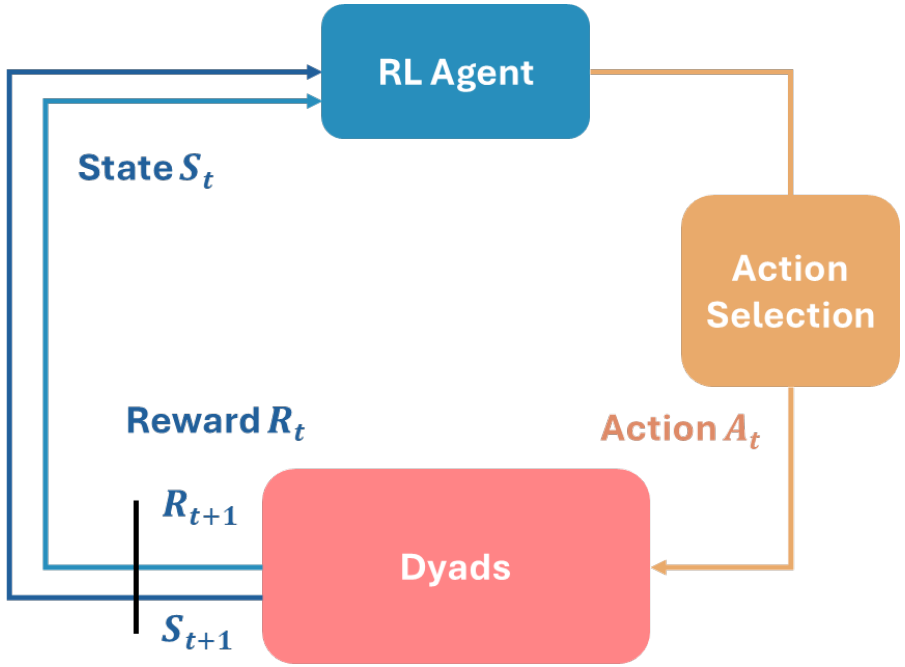
Relationship quality

Self-care strategies

Care partner

Online RL

- **Reinforcement Learning**
  - State S_t
  - Action A_t
  - Reward R_t
- **Goal:** maximize cumulative rewards, $\sum_{t\geq 0} \gamma^t R_t$



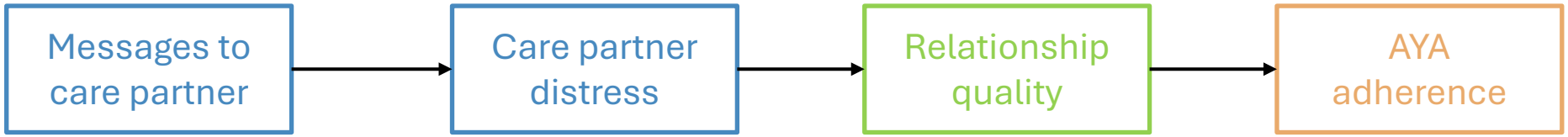
Online RL

- Use streaming data from real-time interactions to update/optimize policy
- "Learning while treating"

RL algorithm is **part of the digital intervention**

Challenges to Online RL

- **Goal:** maximize cumulative rewards (AYA **medication adherence**) over 100 days
- Challenges
  - Multiple interventions (**complex** policy)
  - Interventions impact dyad over **different time-scales**
  - Some interventions only have **distal effect** on rewards



Improve Learning by Incorporating Domain Expertise

- Intervention content is based on **dynamic theories** about how these interventions will impact health outcomes

- **Theory encoded as a causal directed acyclic graph (DAG)**
  - Conditional independence structure

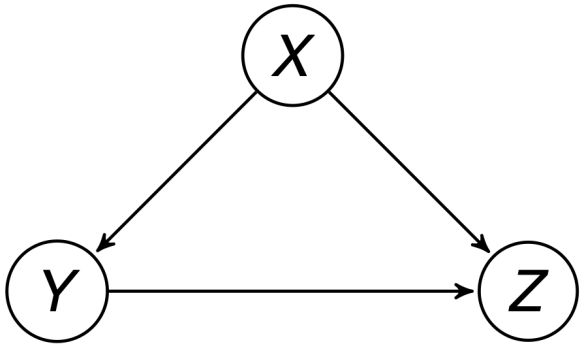


Figure: A Causal DAG Example

How to accelerate RL learning using a DAG?

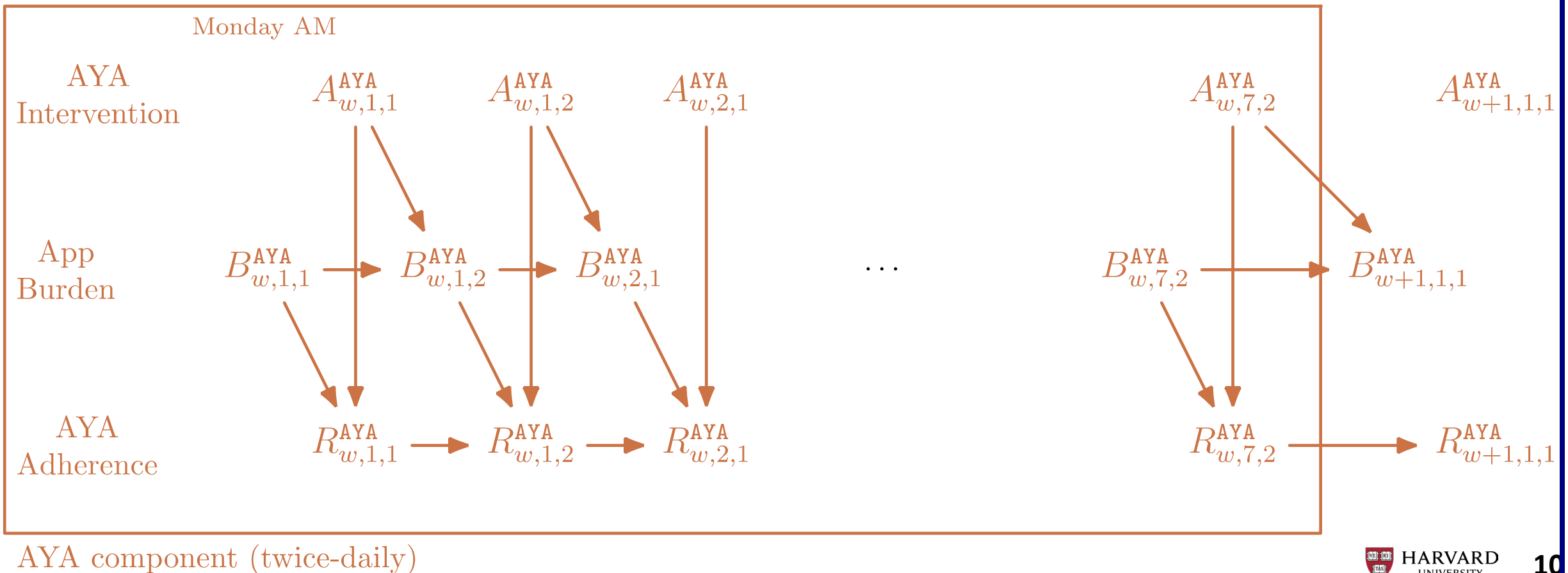
Incorporate Domain Expertise

- Real-world causal relationships are likely very **complex**
  - Direct paths may exist between any two variables

- Scientific team use theory behind the intervention content to reason about the causal pathways that are **most likely detectable** given the **noise and sample size**

- DAGs help RL developer use **these causal pathways** to optimize policy more quickly

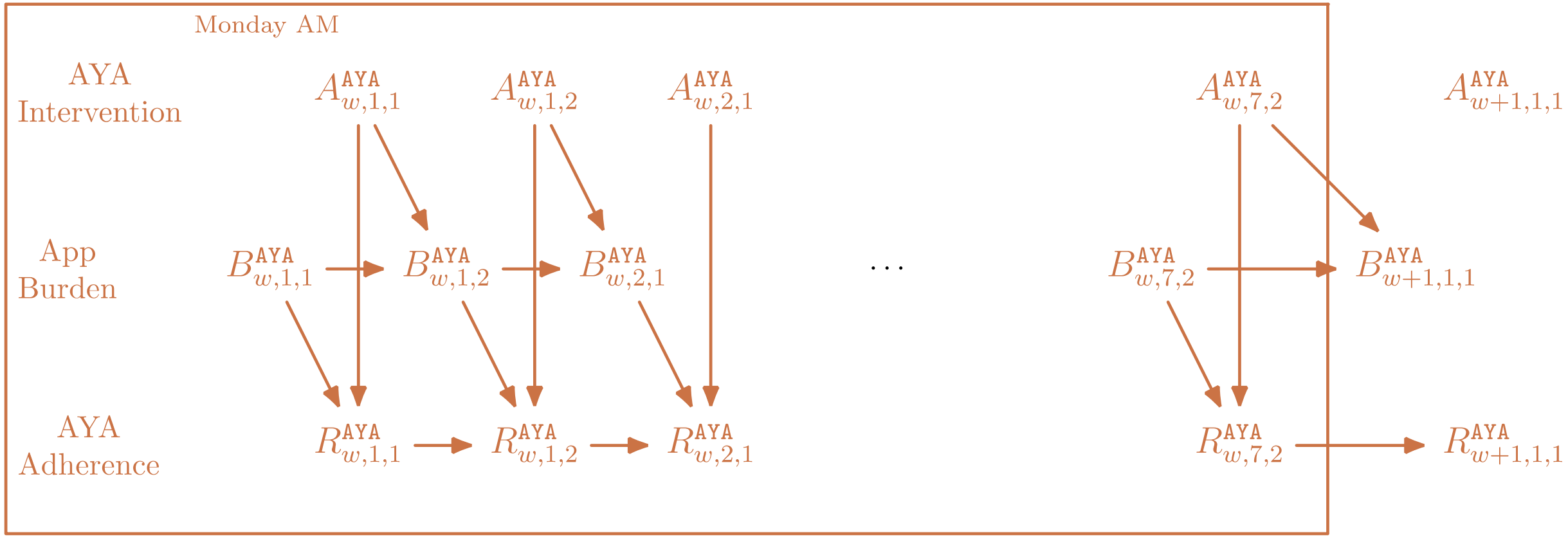
Domain Knowledge as Causal DAG

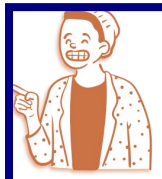


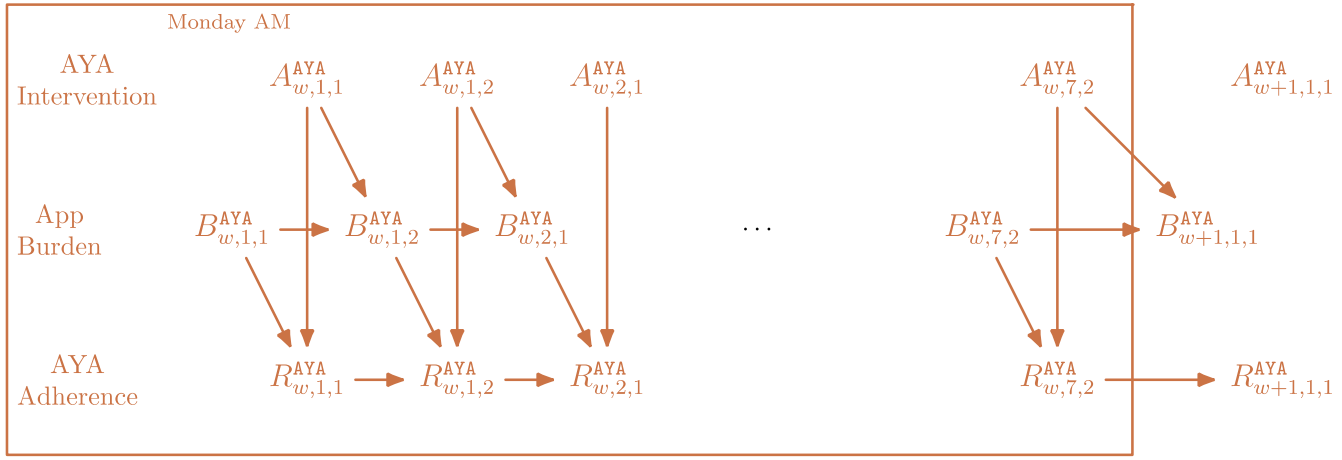
Domain Knowledge as Causal DAG

AYA intervention

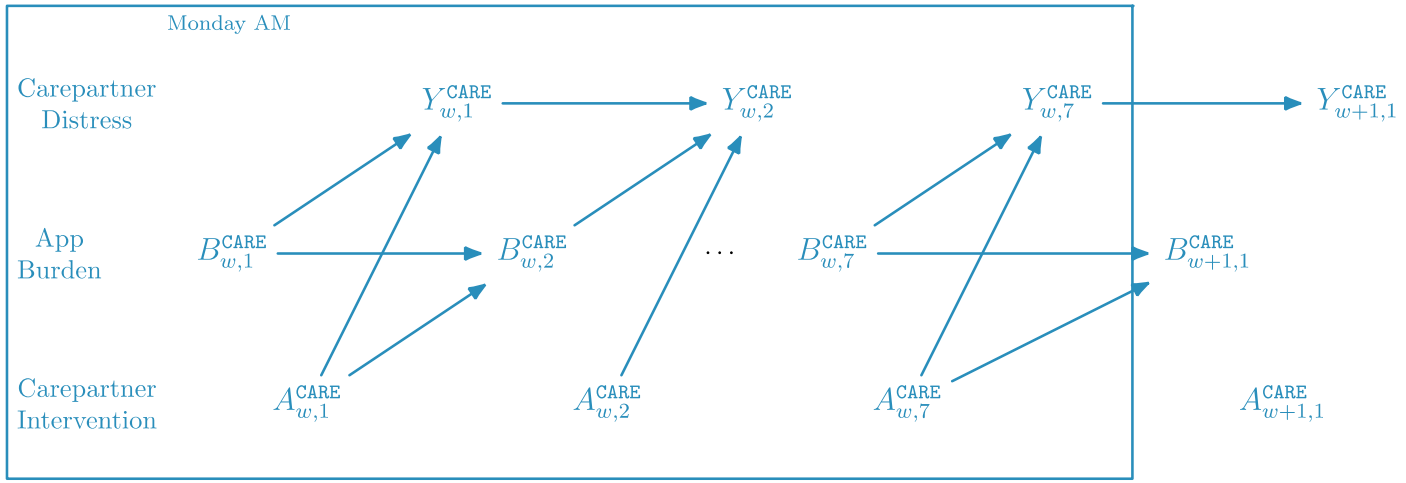
$$A^{\text{AYA}}_{w,d,1} \rightarrow R^{\text{AYA}}_{w,d,1}$$
$$A^{\text{AYA}}_{w,d,1} \rightarrow B^{\text{AYA}}_{w,d,2} \rightarrow R^{\text{AYA}}_{w,d,2}$$



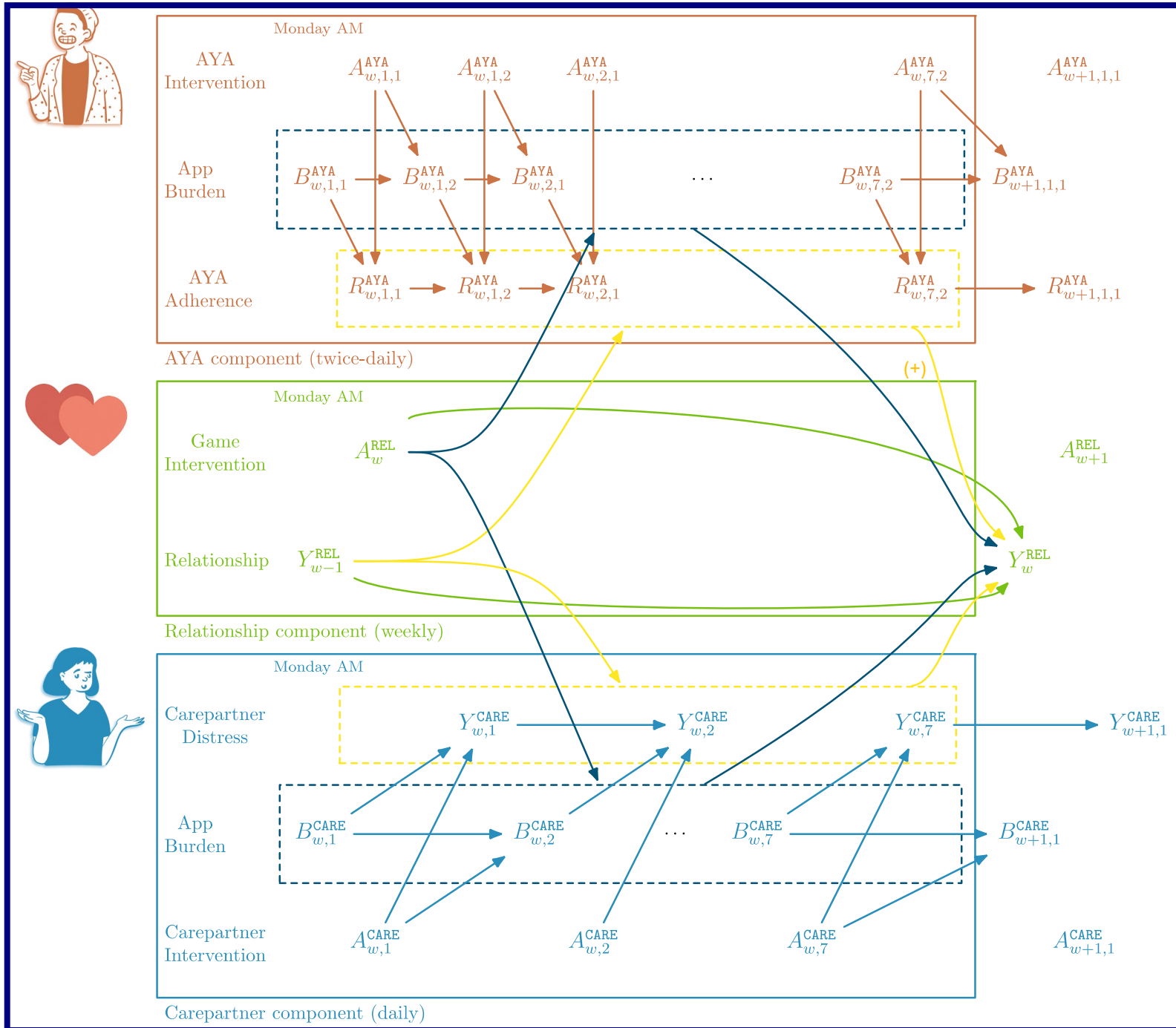


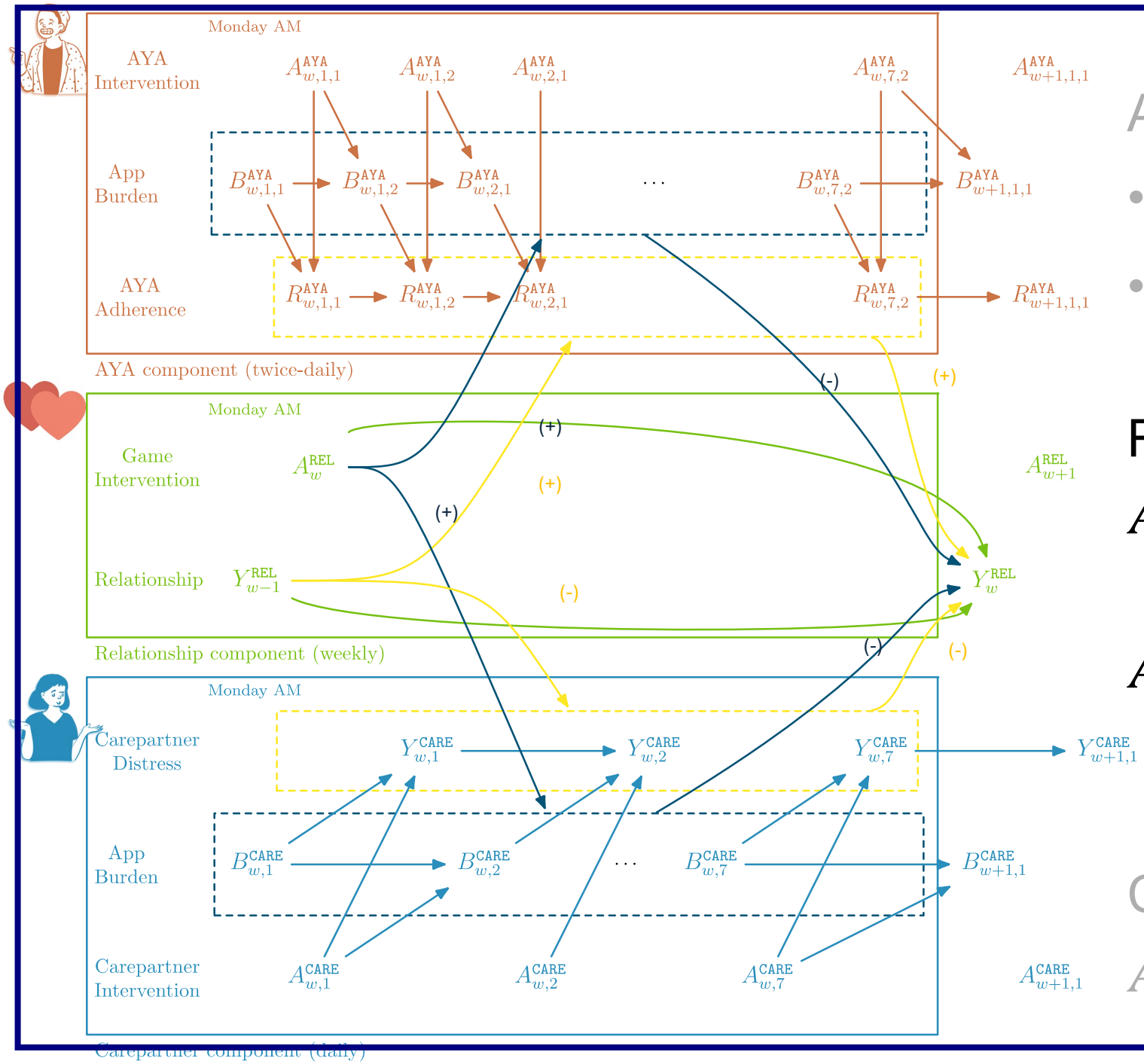


AYA component (twice-daily)



Carepartner component (daily)


Monday AM
AYA Intervention
App Burden
AYA Adherence
AYA component (twice-daily)
(+)
Monday AM
Game Intervention
Relationship
Relationship component (weekly)
Monday AM
Carepartner Distress
App Burden
Carepartner Intervention
Carepartner component (daily)



Some Causal Pathways

AYA intervention

- $A^{\text{AYA}}_{w,d,1} \rightarrow R^{\text{AYA}}_{w,d,1}$
- $A^{\text{AYA}}_{w,d,1} \rightarrow B^{\text{AYA}}_{w,d,2} \rightarrow R^{\text{AYA}}_{w,d,2}$

Relationship intervention

$A^{\text{REL}}_{w} \rightarrow B^{\text{AYA}}_{w,*,*} \rightarrow R^{\text{AYA}}_{w,*,*}$

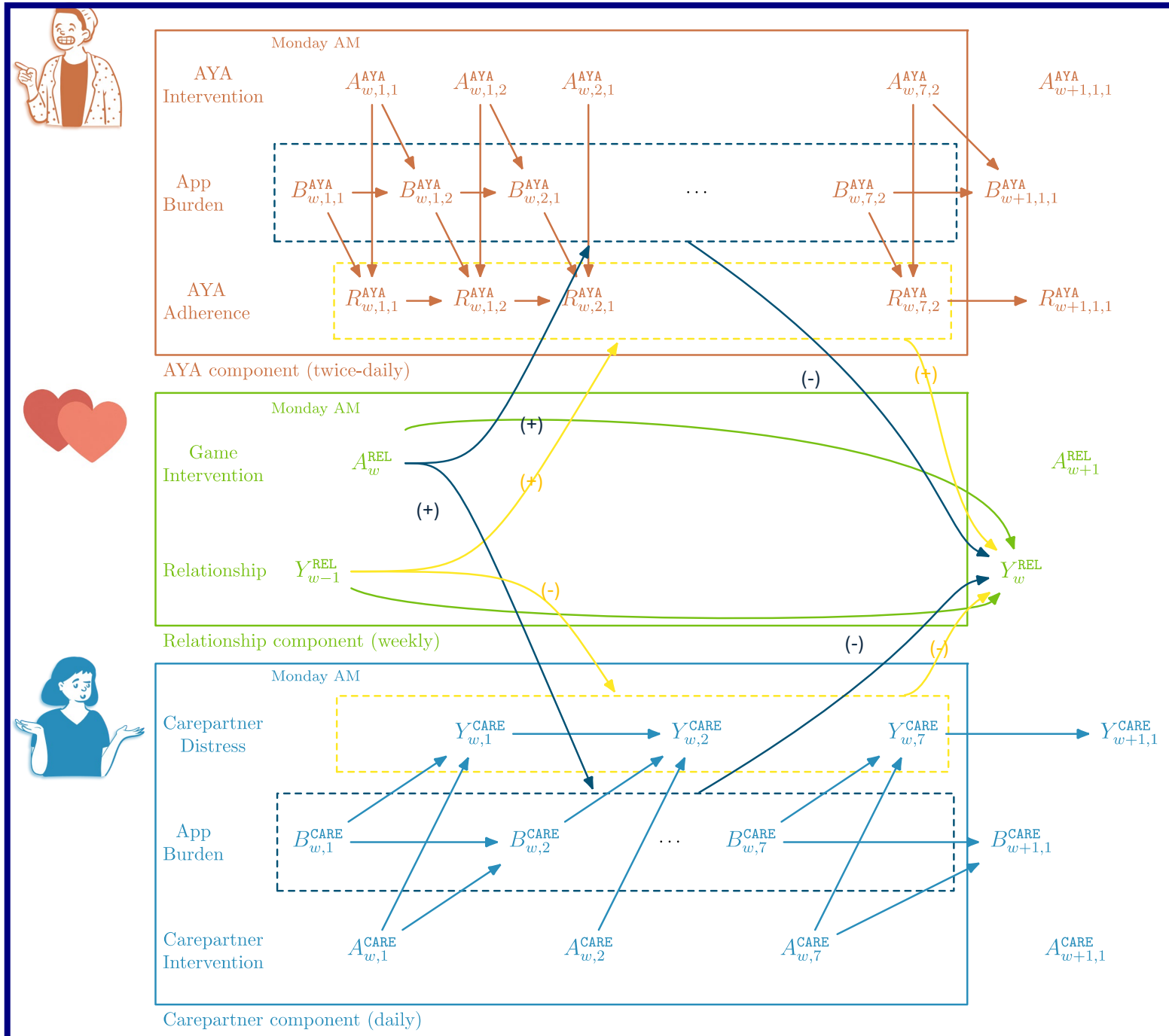
- (burden)

$A^{\text{REL}}_{w} \rightarrow Y^{\text{REL}}_{w} \rightarrow R^{\text{AYA}}_{w+1,*,*}$

- (relationship support)

Care partner intervention

$A^{\text{CARE}}_{w,d} \rightarrow Y^{\text{CARE}}_{w,d} \rightarrow Y^{\text{REL}}_{w} \rightarrow R^{\text{AYA}}_{w+1,*,*}$

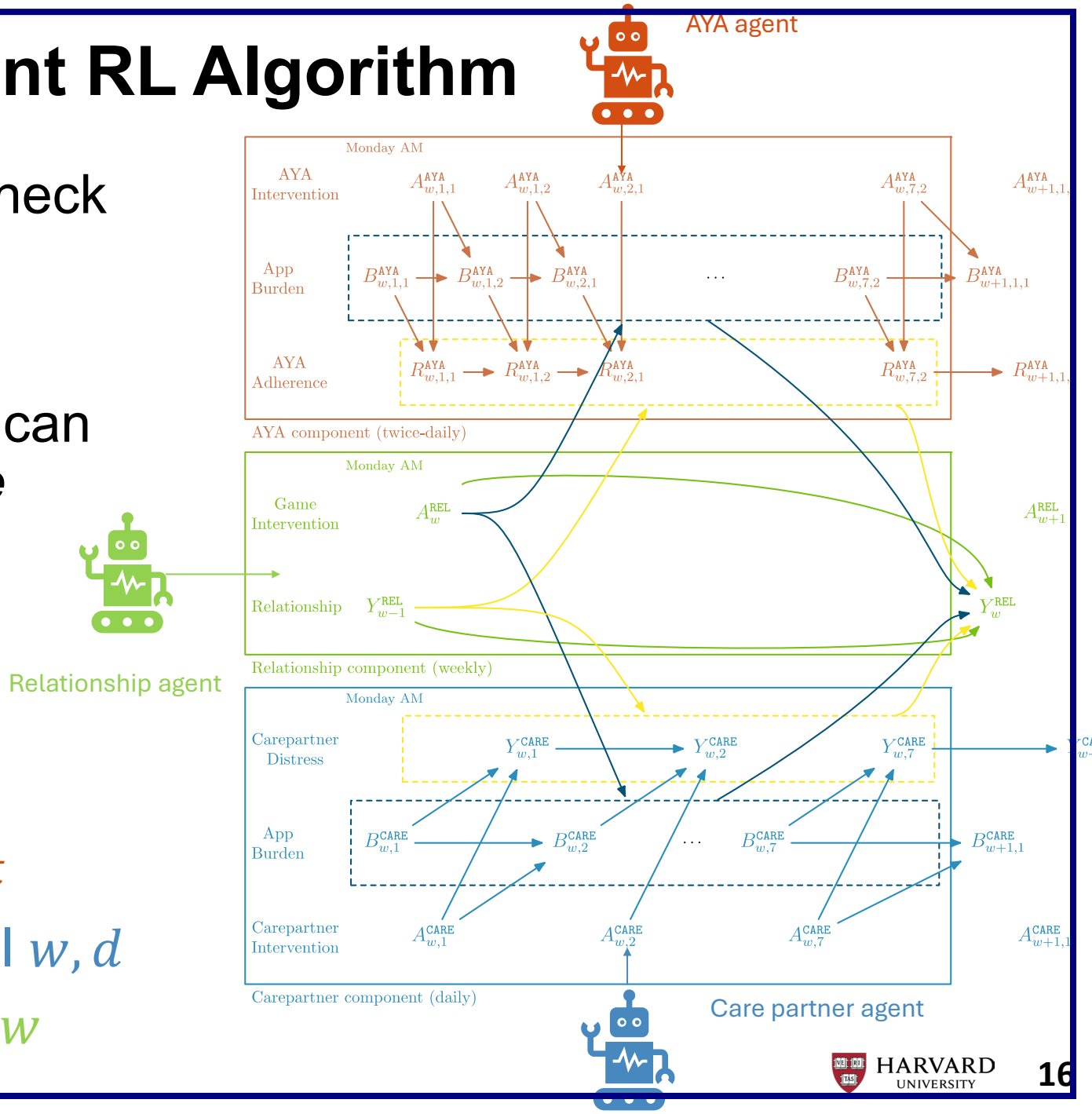


Using the DAG in RL design

- Use DAG to trade bias with variance
  - Identify bottleneck states using DAG
  - State construction
  - Reward construction

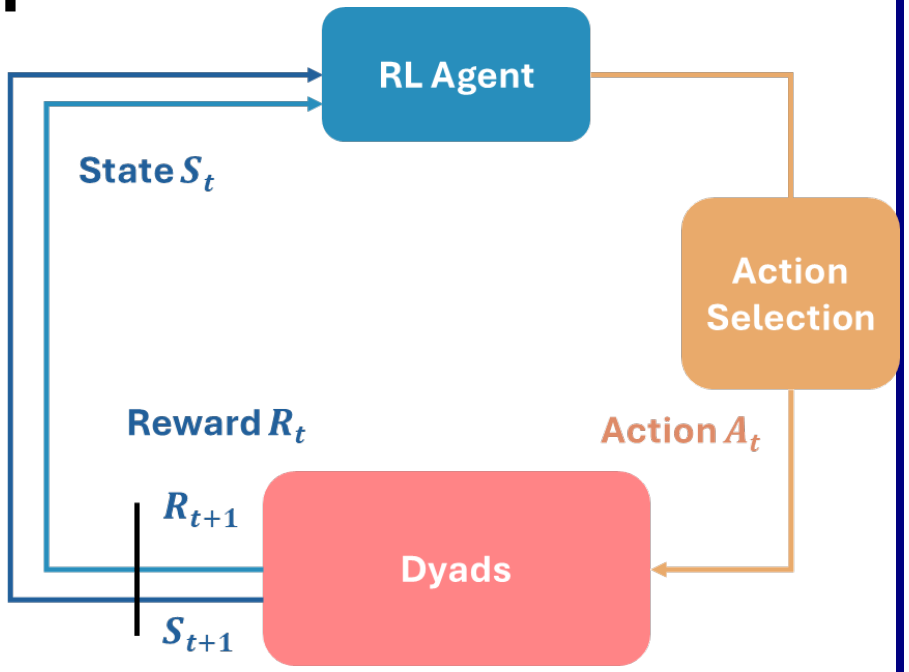
Multi-Agent RL Algorithm

- Relationship component is a bottleneck component
- Conditioned on Y_{w-1}^{REL} and A_w^{REL}, we can *decouple AYA component and care partner component*
- Multi-agent RL
  - AYA agent: $A_{w,d,t}^{\text{AYA}}$ for all w, d, t
  - Care partner agent: $A_{w,d}^{\text{CARE}}$ for all w, d
  - Relationship agent: A_w^{REL} for all w



Base Algorithm

- **Randomized Least Squares Value Iteration + Posterior Sampling**
- **Goal:** maximize $\sum_{t=0}^{\infty} \gamma^t R_t$



- Uses "Bayesian" ideas to estimate $Q(s,a)$

$$Q(s,a) \triangleq E\left[\sum_{t=n}^{\infty} \gamma^{t-n} R_t \middle| S_n = s, A_n = a\right]$$

$$Q(s,a) = E\left[R_n + \gamma \max_{a'} Q(S_{n+1}, a') \middle| S_n = s, A_n = a\right]$$

Base Algorithm: RLSVI

- Advantages:
  - Induce a stochastic policy (action selection probability strictly between 0, 1)
    - Facilitates between-deployments data analysis
  - Bayesian framework used to incorporate existing knowledge as prior
    - Warm-up start/initialization of algorithm

Base Algorithm: RLSVI

Three RLSVI agents:

- AYA agent, Care partner agent, Relationship agent

Hyperparameters for each RLSVI agent

1. Construct states using DAG
2. Select discount rate γ
   - How long the agent looks into the future when optimizing $(1/(1-\gamma))$
3. Construct Algorithm's reward using DAG

1. Agent-Specific States

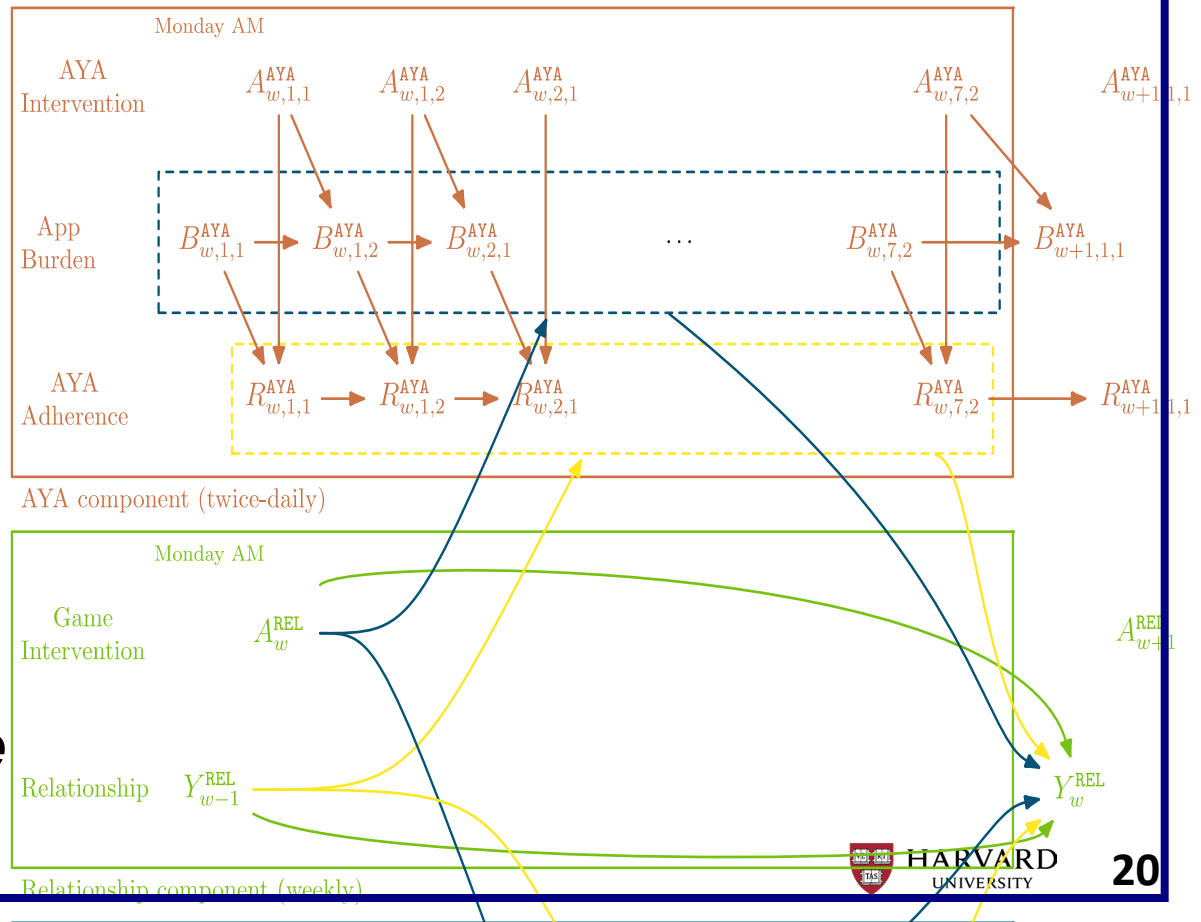
- AYA agent: $\phi_{w,d,2}^{\text{AYA}} = (R_{w,d,1}^{\text{AYA}}, B_{w,d,2}^{\text{AYA}}, Y_{w-1}^{\text{REL}}, A_{w}^{\text{REL}})$

- Care partner agent:

$\phi_{w,d}^{\text{CARE}} = (Y_{w,d-1}^{\text{CARE}}, B_{w,d-1}^{\text{CARE}}, Y_{w-1}^{\text{REL}}, A_{w}^{\text{REL}})$

- Relationship agent:

$\phi_{w}^{\text{REL}} = \left(A_{w-1}^{\text{REL}}, Y_{w-1}^{\text{REL}}\right)$ + (*summary statistics from AYA and care partner)*



2. Agent-Specific Discount Factors

- Discount factor: how much attention should RL pay to future effects?
- Higher discount factor, γ, → more attention to future effects but harder to learn
- Approximate horizon $1/(1-\gamma)$

Maximize:
$\sum_{i\geq 0} \gamma^i R_{t+i}$

- Our decision: All agents have approximate horizon = one week
- AYA agent: $1/(1-\gamma^{\mathrm{AYA}}) = 14$
- Care partner agent: $1/(1-\gamma^{\mathrm{CARE}}) = 7$
- Relationship agent: $\gamma^{\mathrm{REL}} = 0$ (bandit)

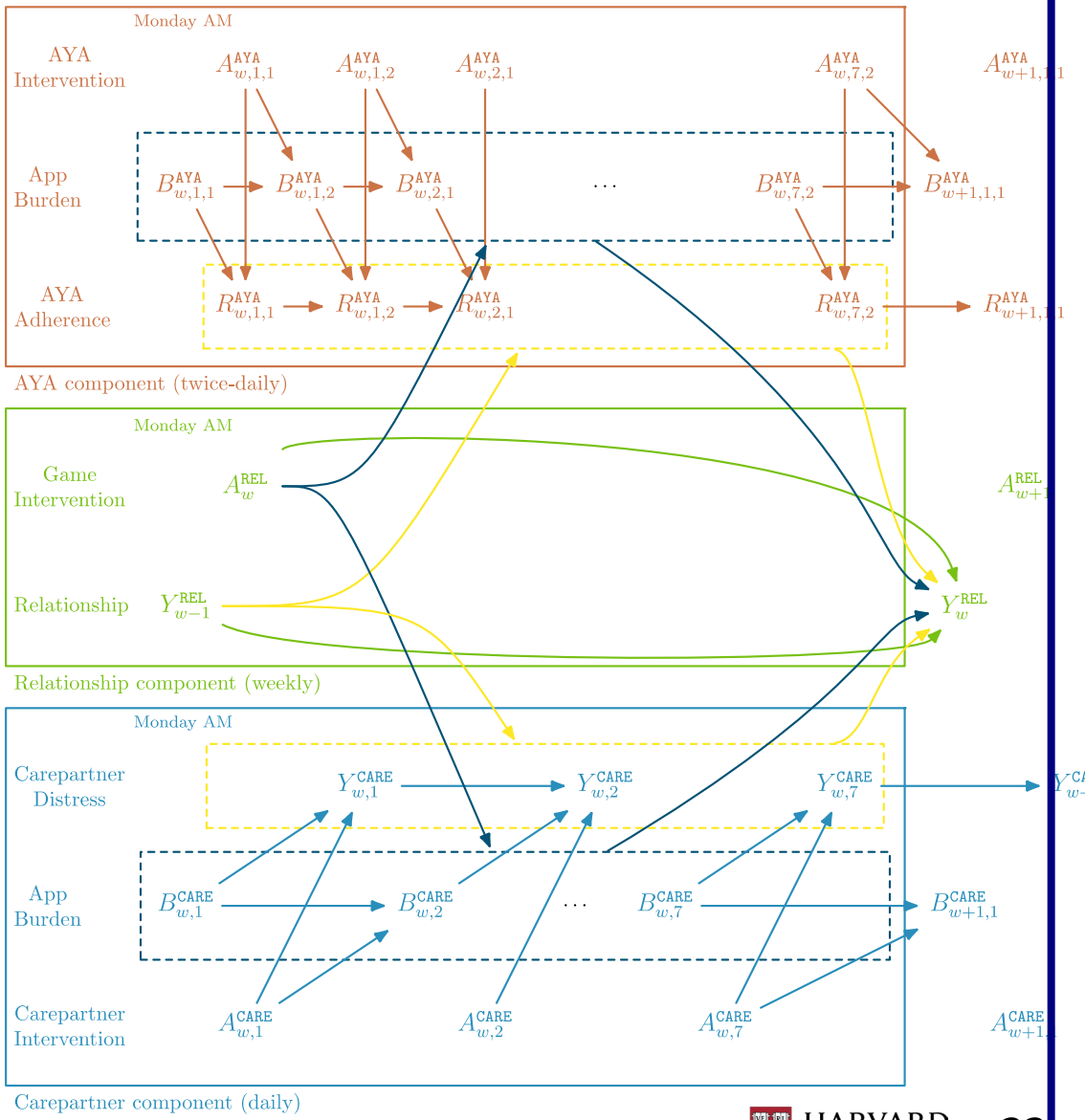
Price of Small Discount Factor, γ

We use small γ to improve learning

Relationship agent

- Bandit algorithm ($\gamma^{\text{REL}} = 0$) does not account for likely positive effect of actions!
  - $A_w^{\text{REL}} \rightarrow B_{w,*,*}^{\text{AYA}} \rightarrow R_{w,*,*}^{\text{AYA}}$ (burden)
  - $A_w^{\text{REL}} \rightarrow Y_w^{\text{REL}} \rightarrow R_{w+1,*,*}^{\text{AYA}}$ (benefit)

Ignoring all delayed effects into future week(s) is problematic!

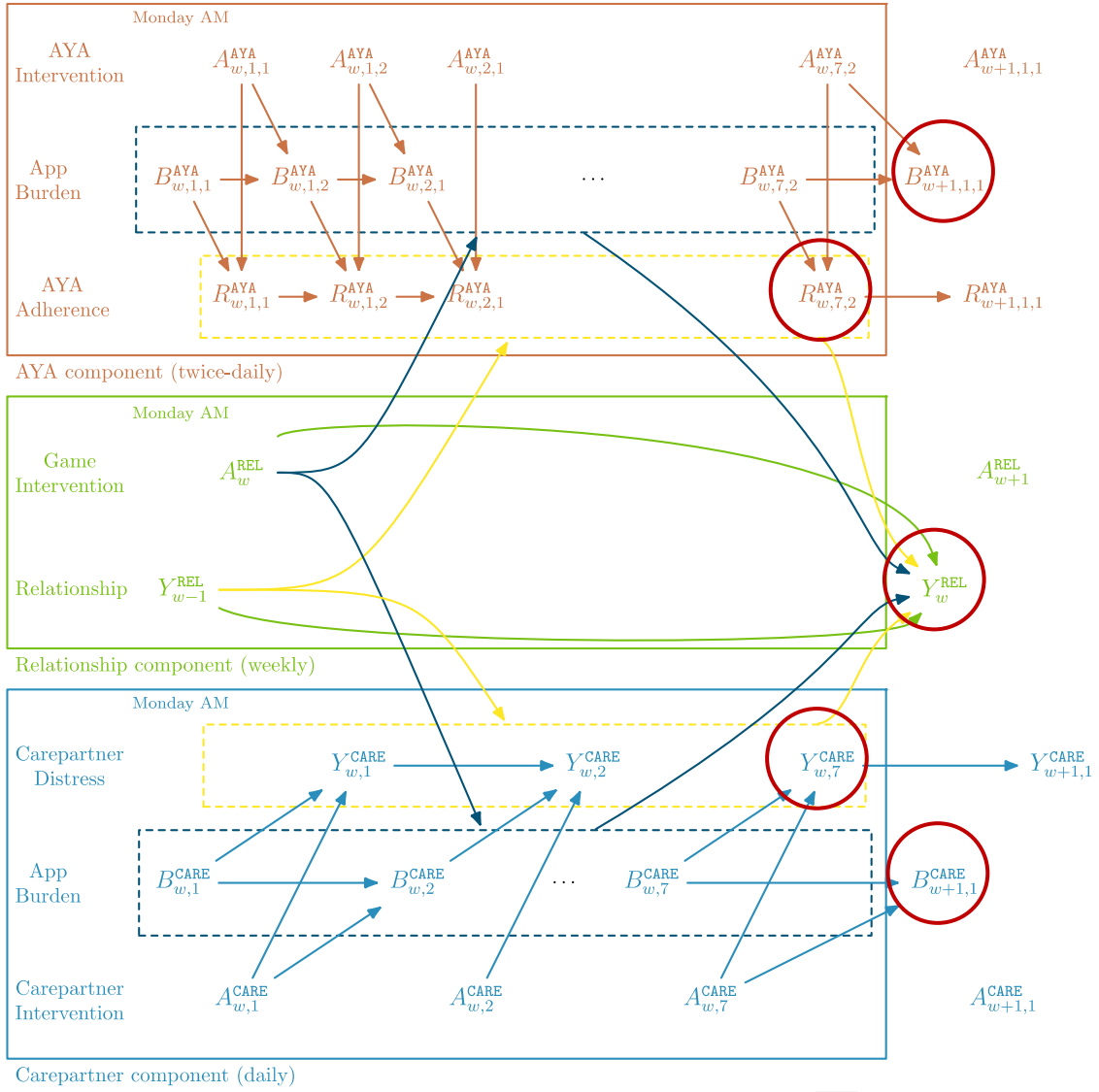


3. Reward Construction

- Use causal DAG

- Identify **mediators** for impact of actions.
  - Relationship agent, A_w^{REL}, on rewards in week $w+1$:
    - $Y_w^{\mathrm{REL}}, B_{w+1,1,1}^{\mathrm{AYA}}, R_{w,7,2}^{\mathrm{AYA}}$

- Use the conditional expectation of next week rewards given these mediators in reward construction



Surrogate Reward for the Relationship Agent

- Action: A_w^{REL}
  - Mediators: $Y_w^{\text{REL}}, B_{w+1,1,1}^{\text{AYA}}, R_{w,7,2}^{\text{AYA}}$

- We estimate the conditional expectation:

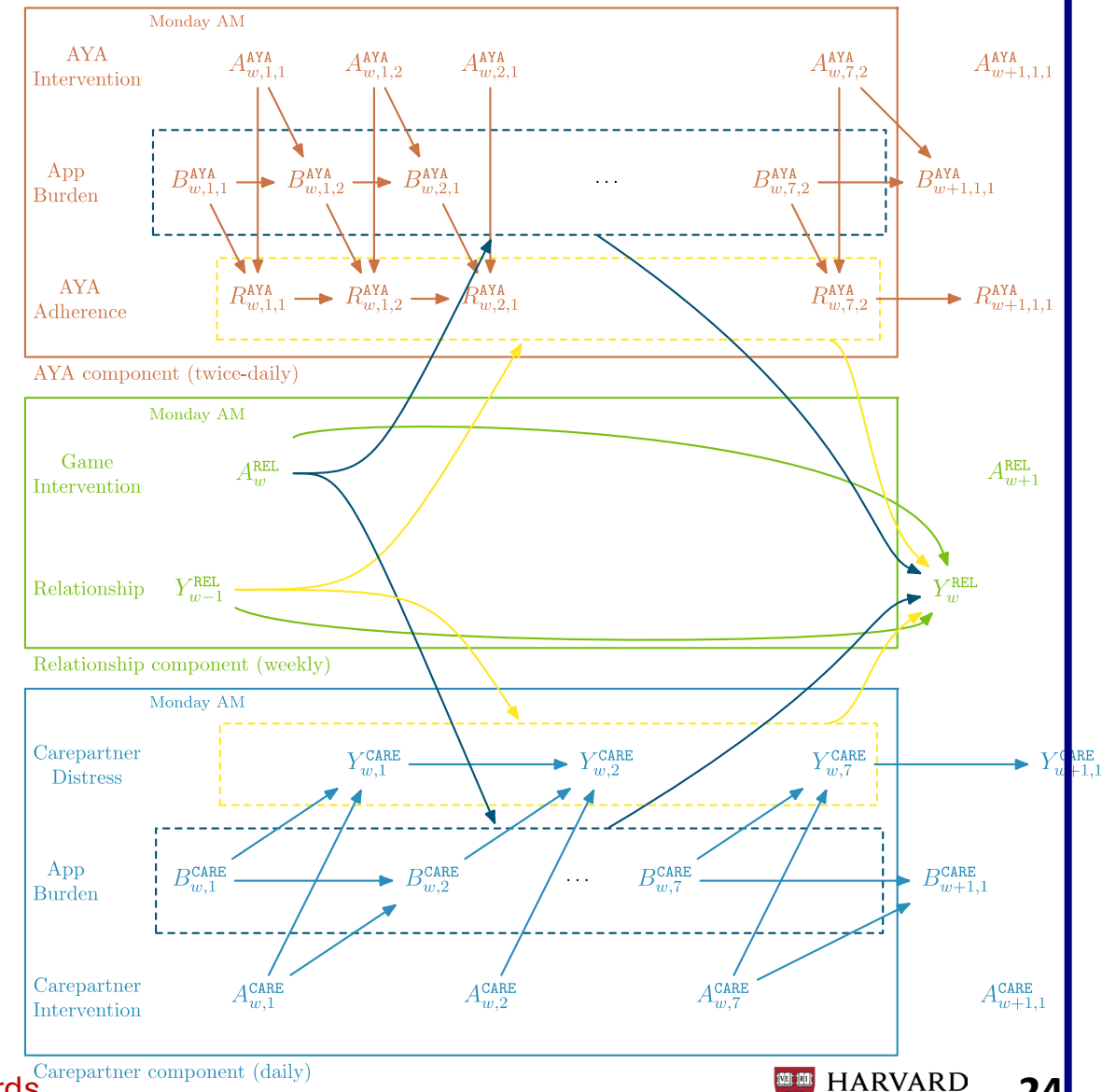
$$f\left(Y_w^{\text{REL}}, B_{w+1,1,1}^{\text{AYA}}, R_{w,7,2}^{\text{AYA}}, A_{w+1}^{\text{REL}}\right) = \mathbb{E}\left[\sum_{d=1,\ldots,7} R_{w+1,d,1}^{\text{AYA}} + R_{w+1,d,2}^{\text{AYA}} \mid Y_w^{\text{REL}}, B_{w+1,1,1}^{\text{AYA}}, R_{w,7,2}^{\text{AYA}}, A_{w+1}^{\text{REL}}\right]$$

- The surrogate reward is a two-step approximation:

$$R_w^{\text{REL}} \triangleq f\left(Y_{w-1}^{\text{REL}}, B_{w,1,1}^{\text{AYA}}, R_{w-1,7,2}^{\text{AYA}}, A_w^{\text{REL}}\right) + \max_a f\left(Y_w^{\text{REL}}, B_{w+1,1,1}^{\text{AYA}}, R_{w,7,2}^{\text{AYA}}, a\right)$$

Current week sum of rewards

Next week sum of rewards

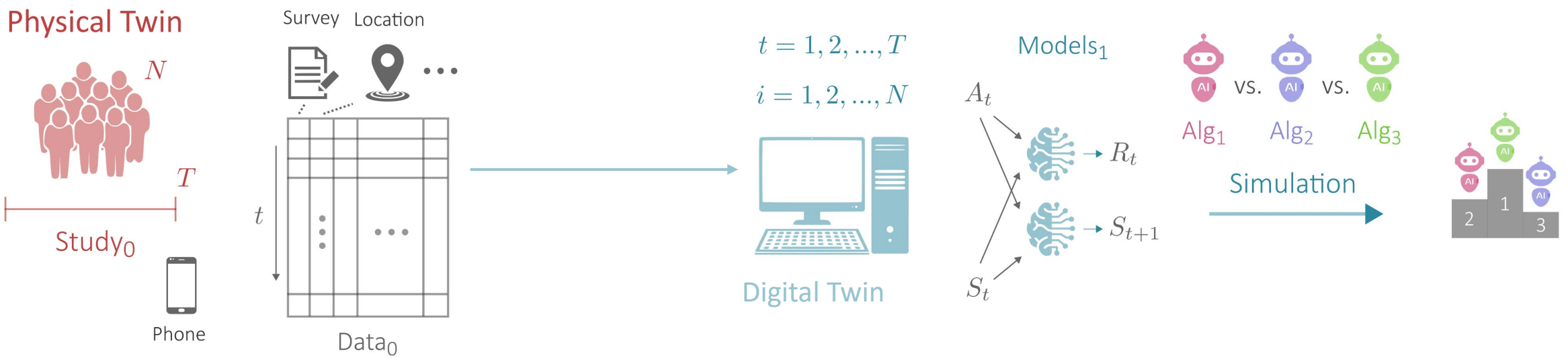


Evaluation of Candidate Algorithms

Build a "digital twin" of the target population

- Based on existing data
- Replicate the expected within person autocorrelation in noise and between person heterogeneity

Roadmap 2.0: existing dataset about cancer patients in dyads

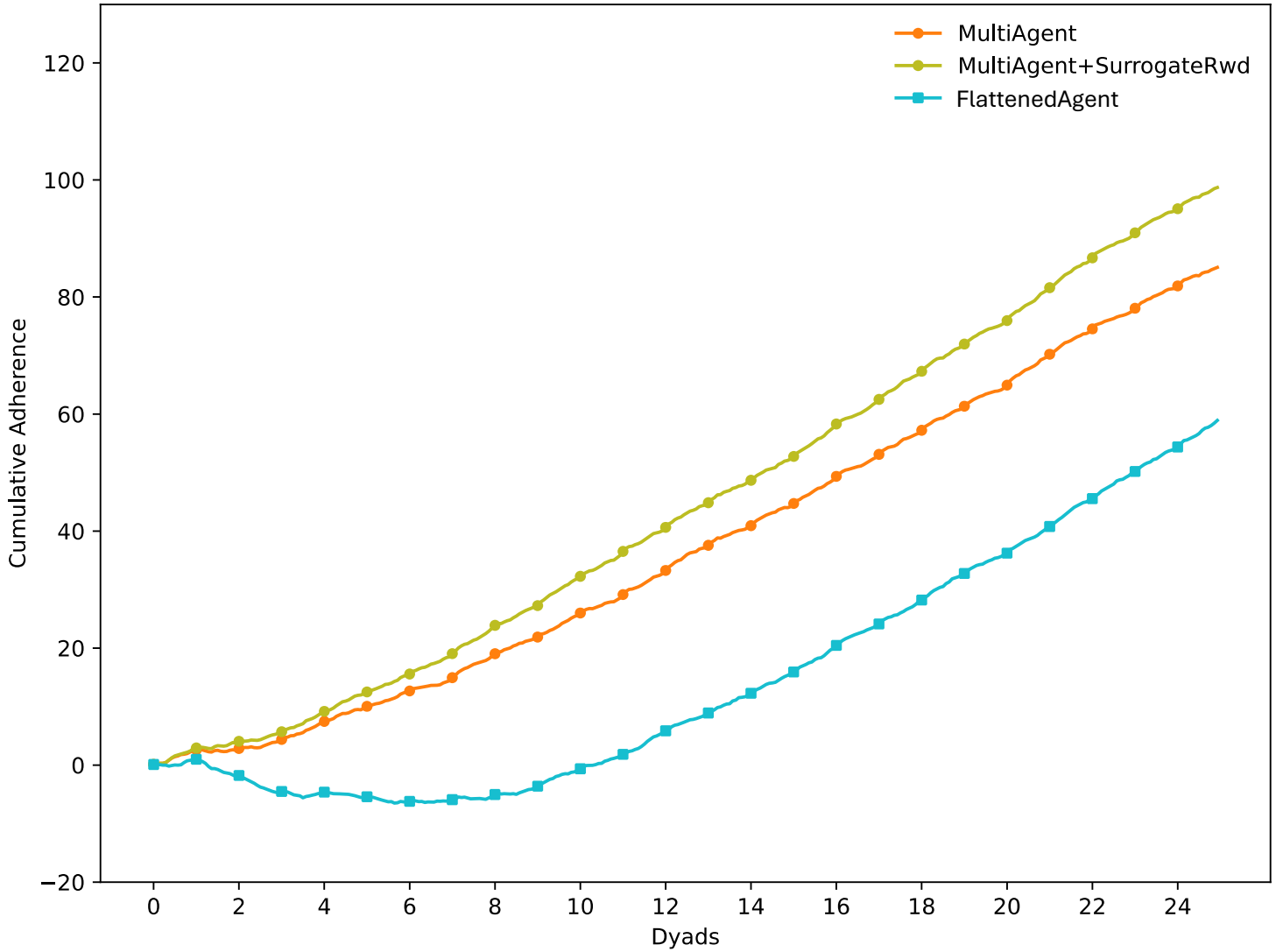


Gazi, A. H., Gao, D., Ghosh, S., Xu, Z., Trella, A., Klasnja, P., & Murphy, S. A. (2025). *Digital twins for just-in-time adaptive interventions (JITAI-Twins): A framework for optimizing and continually improving JITAIs*, Journal of Medical Internet Research. 20/12/2025:72830

Improvement on Cumulative Adherence

Candidate algorithms

1. MultiAgent
2. MultiAgent+SurrogateRwd
3. FlattenedAgent



Q&A

Thank YOU!

Xu Z., Jajal H., Choi S.W., Nahum-Shani, I., Shani G., Psihogios A.M., Hung P, S. Murphy.
Reinforcement Learning on AYA Dyads to Enhance Medication Adherence.
(Appeared in 23rd International Conference on AI in Medicine (AIME-25) Proceedings)

Optimization

- **Digital intervention policy** inputs state and outputs whether to provide and if so what type of digital support to an individual

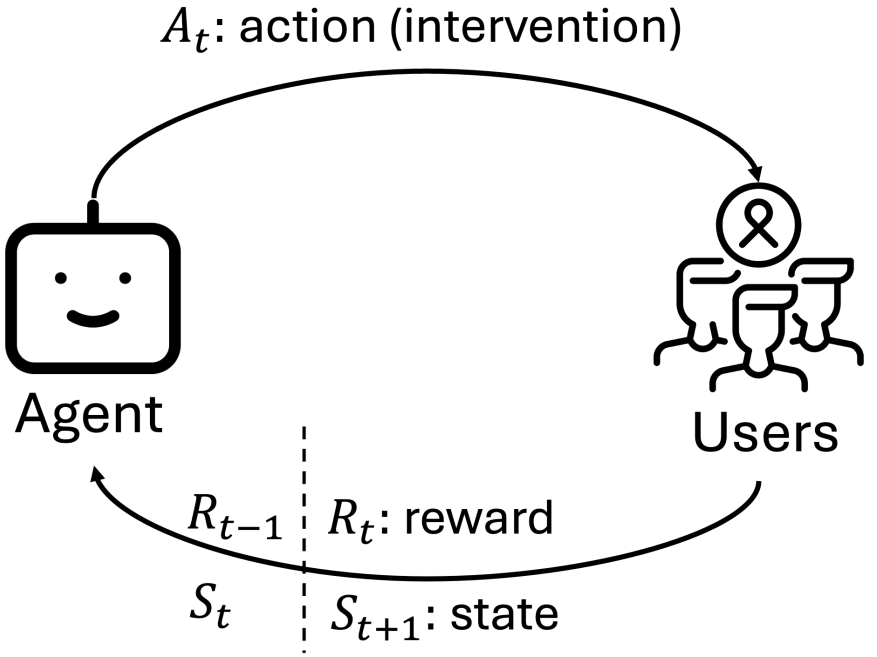
- **Repeated optimization**: In *each* deployment of digital intervention, optimize policy as individual(s) experience the digital intervention.

Why?

- Society, technology (new sensors, operating systems) are changing rapidly resulting in non-stationarity across deployments of the digital intervention.

Why Online RL?

- **Reinforcement Learning**
  - State S_t
  - Action A_t
  - Reward R_t
- **Goal:** maximize cumulative rewards, $\sum_{t=0}^{\infty} \gamma^t R_t$



Online RL

- Optimize policy using streaming data from real-time interactions
- "Learning to treat"

RL algorithm is **part of the digital intervention**

Base Algorithm: RLSVI

$$Q(s,a) = E\left[R_n + \gamma \max_{a'} Q(S_{n+1}, a') \middle| S_n = s, A_n = a\right]$$

- **Randomized Least Square Value Iteration** (RLSVI)[1]: approximate Q-function with a linear model: $Q(s,a)$ by $\phi(s,a)^T\theta$

Algorithm 1 Infinite Horizon RLSVI (Inf-RLSVI)

1: Input: discount factor $\gamma \in \mathbb{R}$, previous dataset $\mathcal{D}_n = (s_i, a_i, r_i)_{i=1}^{n-1} \cup \{s_n\}$, previous perturbation $w \in \mathbb{R}^d$, feature mapping $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, previous parameter $\theta \in \mathbb{R}^d$

2: Generate regression matrix and vector

$$X \leftarrow \begin{bmatrix} \phi(s_1, a_1) \\ \vdots \\ \phi(s_{n-1}, a_{n-1}) \end{bmatrix} \quad y \leftarrow \begin{bmatrix} r_1 + \gamma \max_{\alpha \in \mathcal{A}} \langle \phi(s_2, \alpha), \theta \rangle \\ \vdots \\ r_{n-1} + \gamma \max_{\alpha \in \mathcal{A}} \langle \phi(s_n, \alpha), \theta \rangle \end{bmatrix}$$

3: Estimate value function

$$\bar{\theta} \leftarrow \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^\top X + \lambda I\right)^{-1} X^\top y \quad \mathbf{\Sigma} \leftarrow \left(\frac{1}{\sigma^2}X^\top X + \lambda I\right)^{-1}$$

4: Sample $w' \sim \mathcal{N}(\gamma w, (1-\gamma^2)\mathbf{\Sigma})$ and set $\theta' = \bar{\theta} + w'$

5: **Output:** θ' and w'

[1] Osband, I., Van Roy, B., & Wen, Z. (2016). Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*(pp. 2377-2386). PMLR.

Base Algorithm: RLSVI

$$Q(s,a) = E\left[R_n + \gamma \max_{a'} Q(S_{n+1}, a') \middle| S_n = s, A_n = a\right]$$

- **Randomized Least Square Value Iteration** (RLSVI)[1]: approximate Q-function with a linear model: $Q(s,a)$ by $\phi(s,a)^T\theta$

- Posterior Sampling
  - Select action at time n with probability:

$$P[A_n = a | S_n = s, \mathcal{D}_n] = P\left[\phi(s,a)^T\theta' \geq \max_{a'} \phi(s,a')^T\theta'\right]$$

Algorithm 1 Infinite Horizon RLSVI (Inf-RLSVI)

1: Input: discount factor $\gamma \in \mathbb{R}$, previous dataset $\mathcal{D}_n = (s_i, a_i, r_i)_{i=1}^{n-1} \cup \{s_n\}$, previous perturbation $w \in \mathbb{R}^d$, feature mapping $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, previous parameter $\theta \in \mathbb{R}^d$

2: Generate regression matrix and vector

$$X \leftarrow \begin{bmatrix} \phi(s_1, a_1) \\ \vdots \\ \phi(s_{n-1}, a_{n-1}) \end{bmatrix} \quad y \leftarrow \begin{bmatrix} r_1 + \gamma \max_{\alpha \in \mathcal{A}} \langle \phi(s_2, \alpha), \theta \rangle \\ \vdots \\ r_{n-1} + \gamma \max_{\alpha \in \mathcal{A}} \langle \phi(s_n, \alpha), \theta \rangle \end{bmatrix}$$

3: Estimate value function

$$\bar{\theta} \leftarrow \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^\top X + \lambda I\right)^{-1} X^\top y \quad \mathbf{\Sigma} \leftarrow \left(\frac{1}{\sigma^2}X^\top X + \lambda I\right)^{-1}$$

4: Sample $w' \sim \mathcal{N}(\gamma w, (1-\gamma^2)\mathbf{\Sigma})$ and set $\theta' = \bar{\theta} + w'$

5: **Output:** θ' and w'

"posterior sampling": sample temporally correlated noise to generate randomized weights θ'

[1] Osband, I., Van Roy, B., & Wen, Z. (2016). Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*(pp. 2377-2386). PMLR.

Base Algorithm: RLSVI

- Advantages:
  - Induce a stochastic policy (action selection probability strictly between 0, 1)
    - Facilitates between-deployments data analysis
  - Bayesian framework uses to incorporate existing knowledge as prior
    - Warm-up start/initialization of algorithm

Surrogate Reward for the Care Partner Agent

- The surrogate reward is constructed through two layers
- **Layer 1:**
  - Mediator Y_w^{REL} (believed to have positive effects)
  - $R_{w,d}^{\mathrm{CARE}} = \mathbb{I}\{d = 7\} Y_w^{\mathrm{REL}}$
  - Sparse reward is still hard to learn
- **Layer 2:**
  - Attribute sparse reward into different steps
  - Fit a linear regression to estimate

Context

Outcome of $A_{w,d}^{\mathrm{CARE}}$

$$g\left(Y_{w,d}^{\mathrm{CARE}}, B_{w,d+1}^{\mathrm{CARE}}, Y_{w-1}^{\mathrm{REL}}, A_w^{\mathrm{REL}}, A_{w,d}^{\mathrm{CARE}}\right)$$

$$= \mathbb{E}\left[Y_w^{\mathrm{REL}} \mid Y_{w,d}^{\mathrm{CARE}}, B_{w,d+1}^{\mathrm{CARE}}, Y_{w-1}^{\mathrm{REL}}, A_w^{\mathrm{REL}}, A_{w,d}^{\mathrm{CARE}}\right]$$

- $R_{w,d}^{\mathrm{CARE}} = g\left(Y_{w,d}^{\mathrm{CARE}}, B_{w,d+1}^{\mathrm{CARE}}, Y_{w-1}^{\mathrm{REL}}, A_w^{\mathrm{REL}}, A_{w,d}^{\mathrm{CARE}}\right)$