Improving Embodiment with Reinforcement Learning in Virtual Reality

Final Presentation
Name: Yawen Hou
Supervisor: Thibault Porssut, Ronan Boulic

Goal of the experiment

Dynamically & rapidly find the maximal <u>distortion</u> threshold for each subject

- Need an **robust, online, adaptive and rapidly converging algorithm**

But what is a distortion ?

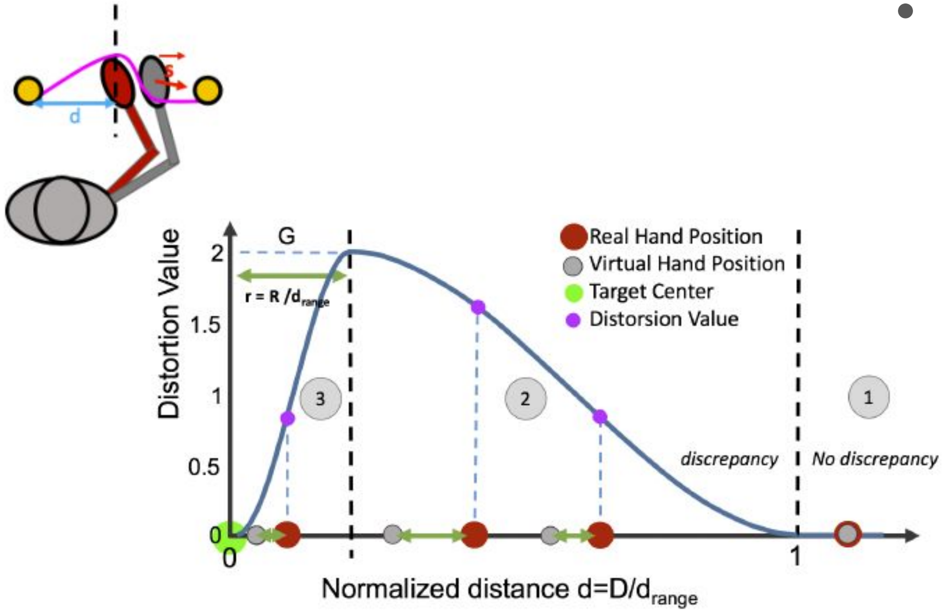
Embodiment in VR

Sense of embodiment (Kilteni, 2012):

- **Self-location:** first-person view
- **Agency:** motor activity control
- **Body ownership:** self-attribution

Break any of these ⇒ **Break In Embodiment (BIE)**

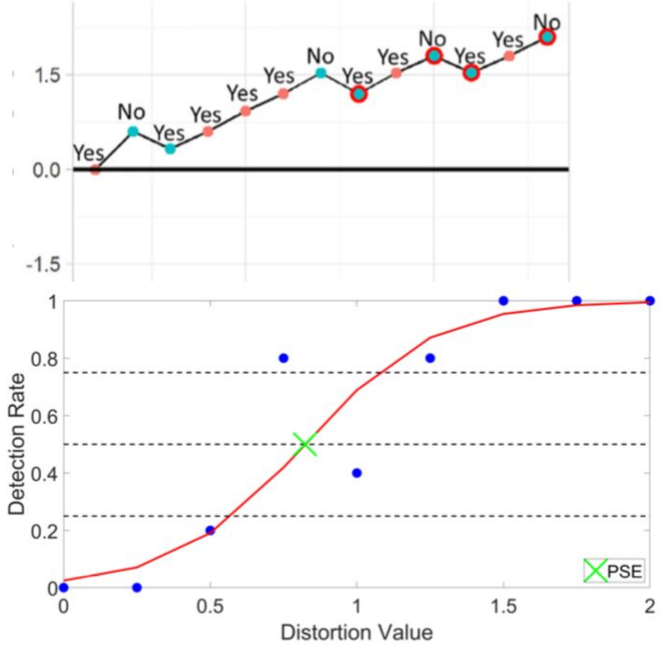
Distortion in VR



- Burns(2006):
  - First to study distortion in VR to study 2 types of discrepancies (interpenetration vs position)
- Bovet, Debarba (2018):
  - Avoid interpenetration of the virtual body while the subject receives a passive haptic feedback from their real body
  - Hindering / helping distortion force while moving towards a static target ⇒ threshold

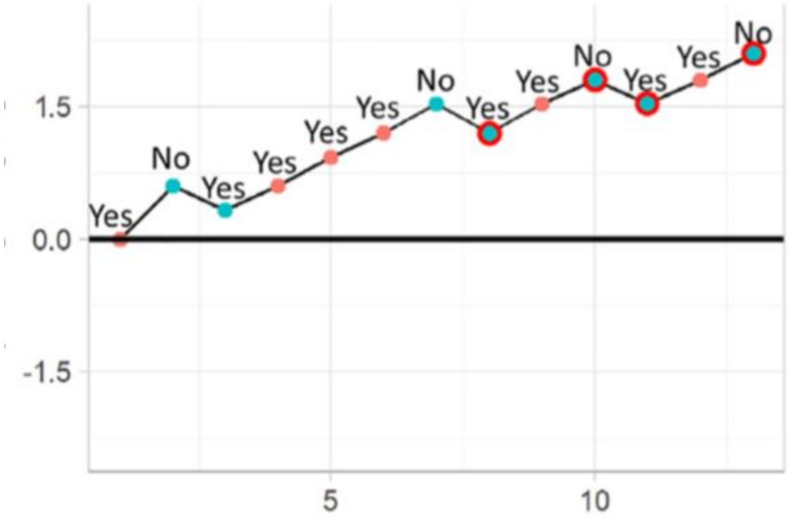
Porssut (2019):
  - Level of tolerance for distortion
  - Moving target
  - Distortion used for the current experience
  - Attraction well: first attract the hand to the outer boundary, then attraction diminishes to 0
  - G ⇒ Parameter to vary for finding the threshold

Finding the Distortion Threshold



- High inter-subject variability of distortion threshold depending on subject’s past experience
- Two previous methods to find the converging value:
    - Staircase (Bovet 2018, Debarba 2018)
        - Online
        - 80 iterations to terminate
    - Point of Subjective Equality (PSE) (Porssut 2019)
        - Offline
        - 45 iterations to gather data

Staircase



- 4 staircases in parallel
- Staircases presented in random order at each round
- Converges: 7 turns in direction
- Distortion threshold: mean of last 4 turns in direction
- Termination: 20 trials

Sidney Bovet, Henrique Galvan Debarba, Bruno Herbelin, Eray Molla, and Ronan Boulic. The critical role of self-contact for embodiment in virtual reality. IEEE Transactions on Visualization and Computer Graphics, 24(4):1428–1436, April 2018

Staircase Limitations

- Not robust
- Not conservative enough
- Not adaptive
- Moderate possibility of non convergence
- Need to run multiple staircases in parallel for the same subject to avoid habituation, but they do not always converge to the same thresholds

Evaluation criteria

Given a threshold T found by an algorithm for a particular subject, we calculate:

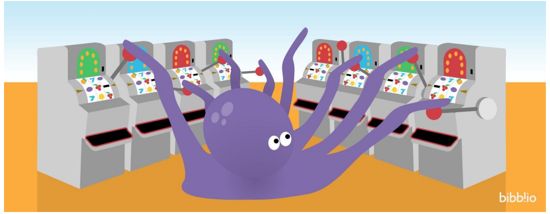
- **Robustness :** $R(\%) = \frac{\text{nb trials gain} <= \text{T AND subject experiences BIE}}{\text{total nb of trials during the experiment}}$
- **Conservativeness:** $C(\%) = \frac{\text{nb trials gain} > \text{T AND no BIE}}{\text{total nb of trials during the experiment}}$
- **Convergence speed:** No of trials needed for the algorithm to converge (each trial takes 6s)
- **Adaptivity:** Ability of the algorithm to update the threshold

Hypotheses

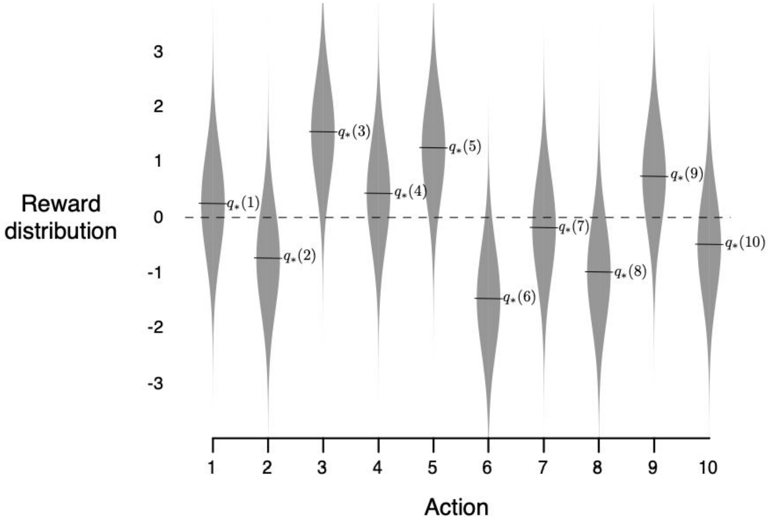
- H1: UCB algorithm is more **robust** than the staircase
- H2 : UCB algorithm is more conservative than the staircase
- H3: UCB algorithm converges in less iterations than the staircase
- H4: UCB algorithm is more adaptive than the staircase

⇒ **Robustness is our most important criteria to avoid a Break In Embodiment (BIE).**

Non-stationary multi-armed bandit



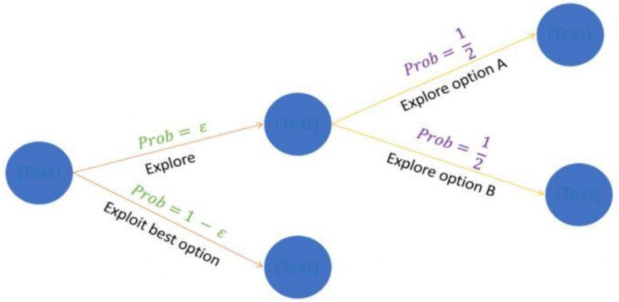
"K slot machines available, at each time, choose to pull the arm of one slot machine and get a immediate reward."



10 armed bandit. Action 3 is the best.

- Action: Machine to choose
  - Discrete distortion values to choose
  - [0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 4, 5, 7, 10]
- Reward: +ve / -ve depending on the reaction of the subject
  - +ve : no BIE;   -ve: BIE
  - Abs(reward) = distortion gain
- **Goal: Maximize total rewards**
- Non-stationary: There is always a best machine, but the choice of the best machine can change over time
- Q-table: Estimate of the winning probabilities of each machine

ϵ-Greedy and Upper Confidence Bounds (UCB)



- **ϵ-Greedy**: For ϵ% of the time, explore a random action. For (1-ϵ)% of the time, exploit the action chosen by UCB

- **UCB**: choose the action among those that have been the least explored in the past

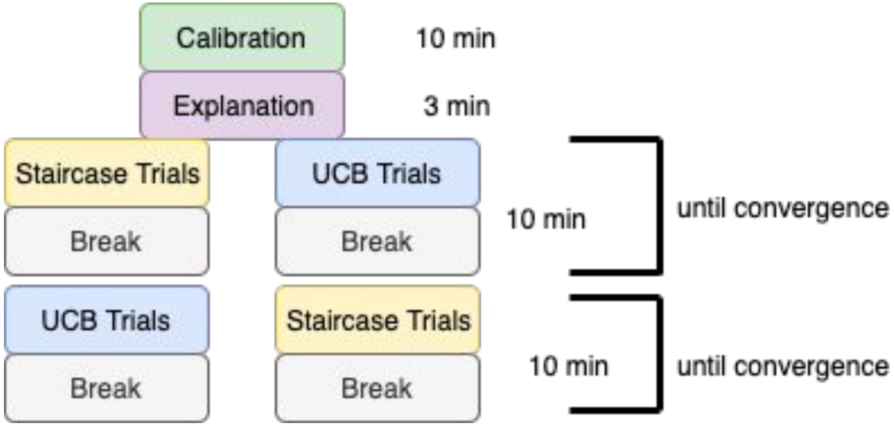
$$A_n \doteq argmax_a \left(Q_n(a) + c\sqrt{\frac{log(n)}{k_n(a)}} \right)$$

- **Q(a)** estimate how good an action a is
- Non-stationary update: $Q_{t+1} = Q_t + \alpha[R_t - Q_t]$
- **Convergence** : When the action with highest Q-value doesn't change for 15 consecutive iterations
- **Terminate**: 100 trials

Hypotheses

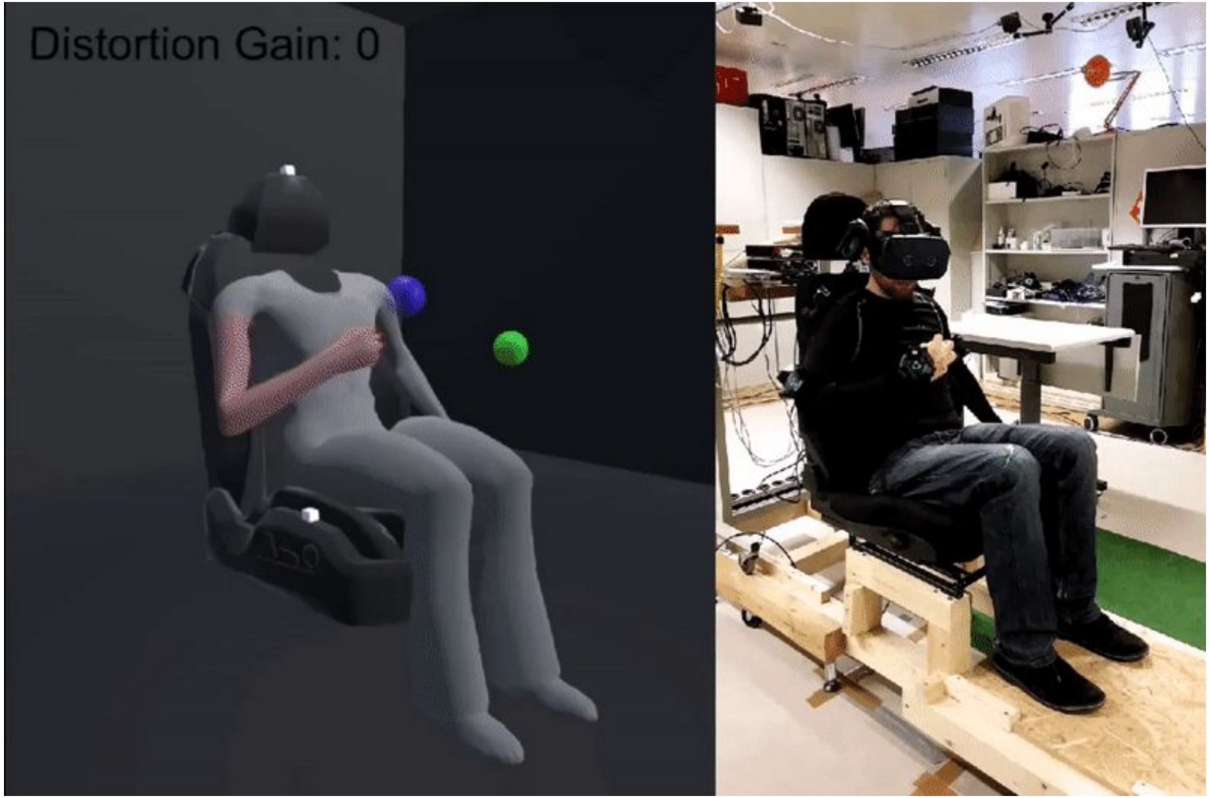
- H1: UCB algorithm is more robust than the staircase
- H2 : UCB algorithm is more conservative than the staircase
- H3: UCB algorithm converges in less iterations than the staircase
- H4: UCB algorithm is more adaptive then the staircase

Experiment protocol



- Pilot study of 5 participants
- 22 subjects (1 data excluded due to technical issue)
    - Aged from 18-25
    - 4 females, 17 males
    - Most of them do not have extensive experience in VR

Experiment Material

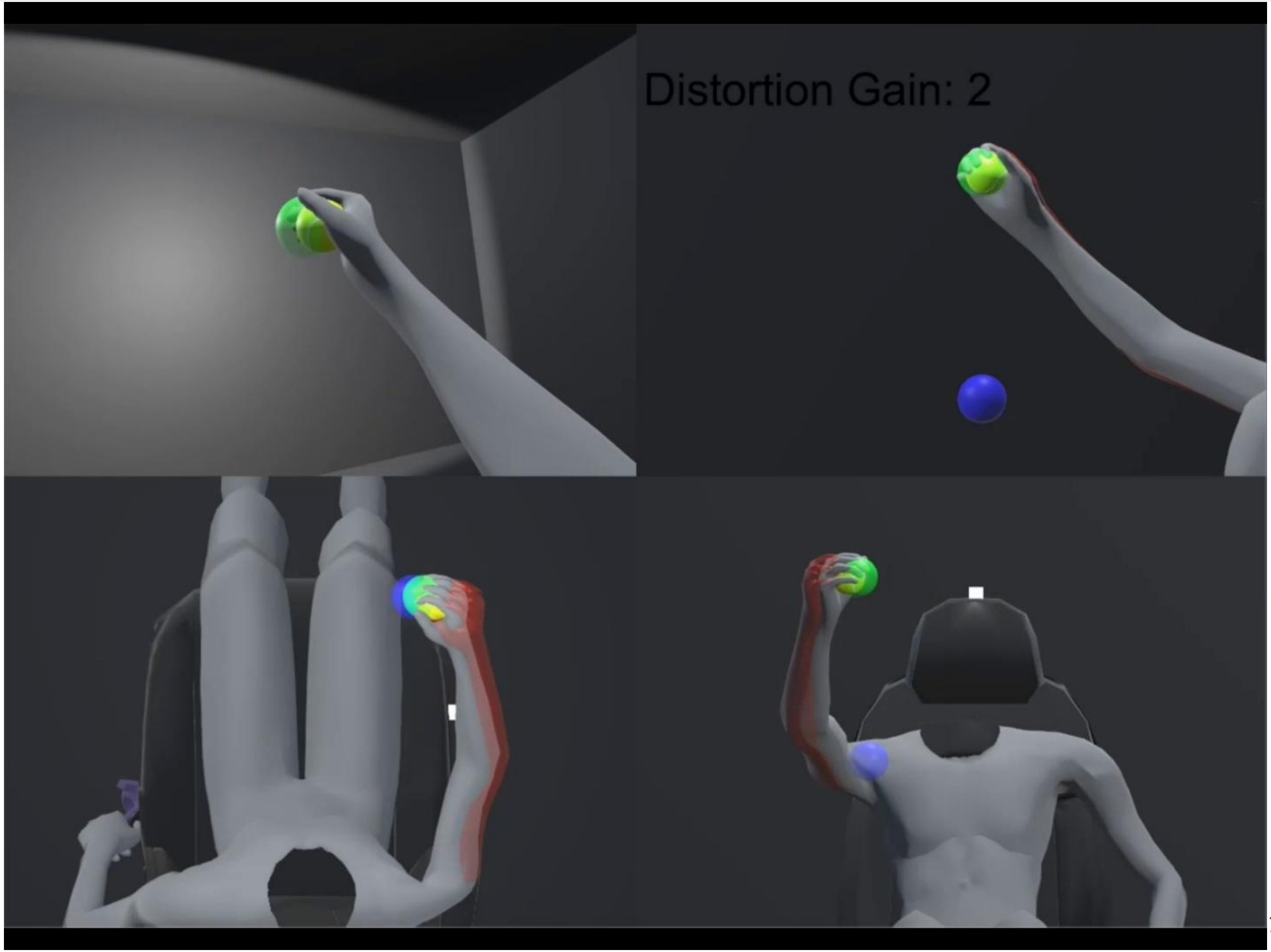


- 1 Vive Pro Eye
- 8 HTC Vive trackers (v2)
- 1 Vive controller
- 1 Bose QuietComfort 35 wireless headphones (not in image)
- 1 tennis ball

Task

- 2 targets
- 1st: Blue sphere
- 2nd (moving): Green sphere

⇒ 2 phase movement



Statistical analysis

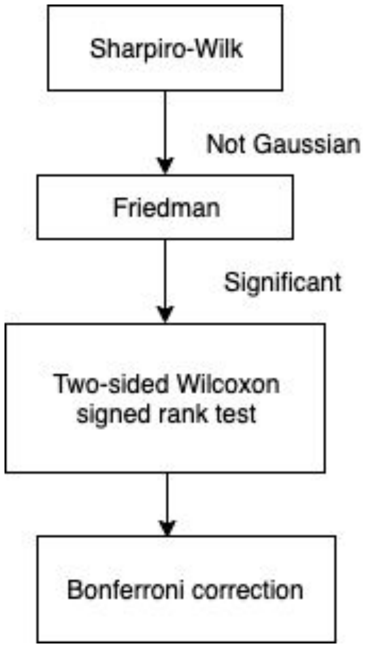
- Comparisons **within** subjects
- 1 independent factor: distortion gain
- Discarded staircases that did not converge

Analysis with 4 levels:

1. Most robust staircase: staircase with smallest R(%)
2. Least conservative staircase: staircase with smallest C(%)
3. In-middle staircase: staircase with smallest absolute difference (R-C)%
4. UCB value

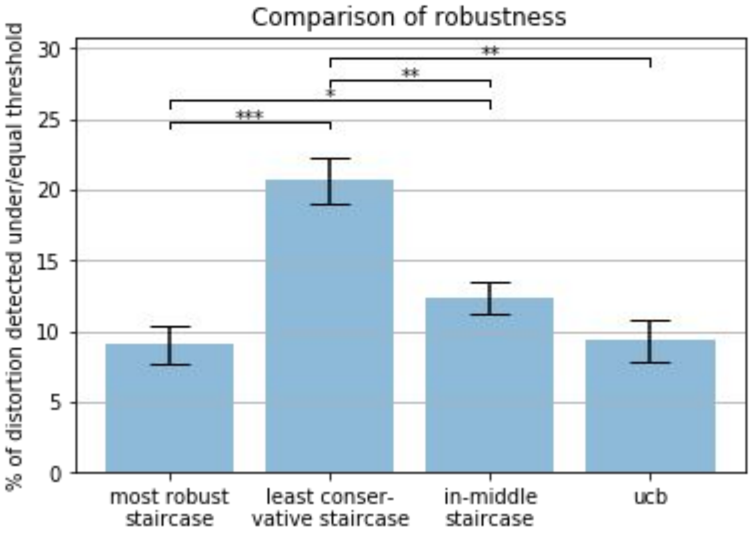
⇒ Robustness: R = % of time subject **sees** distortion when they **should not** (under threshold)

⇒ Conservativeness: C = % of time subject **doesn't see** distortion when they **should** (above threshold)



$$R(\%) = \frac{\text{nb trials gain} <= \text{T AND subject experiences BIE}}{\text{total nb of trials during the experiment}}$$

Criterion 1 - Robustness



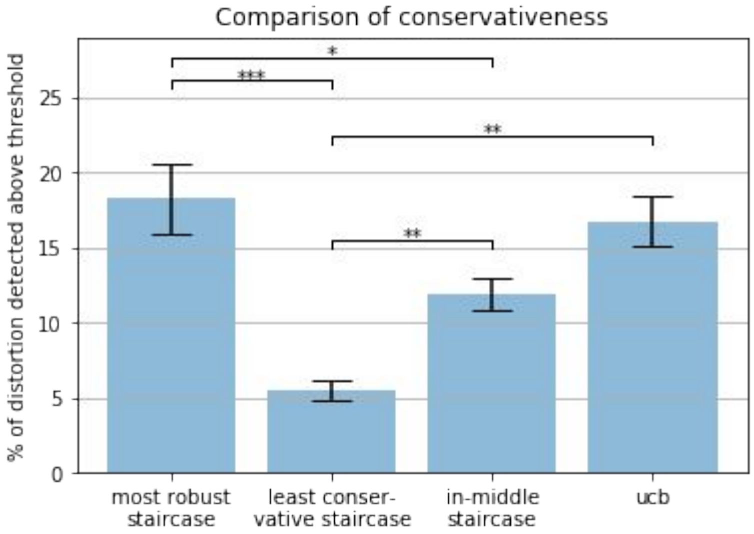
- Robustness: % of time subject notices the distortion when they shouldn't.
- UCB - Most robust staircase
    - No significant difference
- UCB - Least conservative staircase:
    - Significant difference
- UCB - In-Middle staircase:
    - No significant difference

⇒ UCB is as robust as the most robust staircase.

⇒ H1 verified for least robust staircase.

$$C(\%) = \frac{\text{nb trials gain} > \text{T AND no BIE}}{\text{total nb of trials during the experiment}}$$

Criterion 2 - Conservativeness

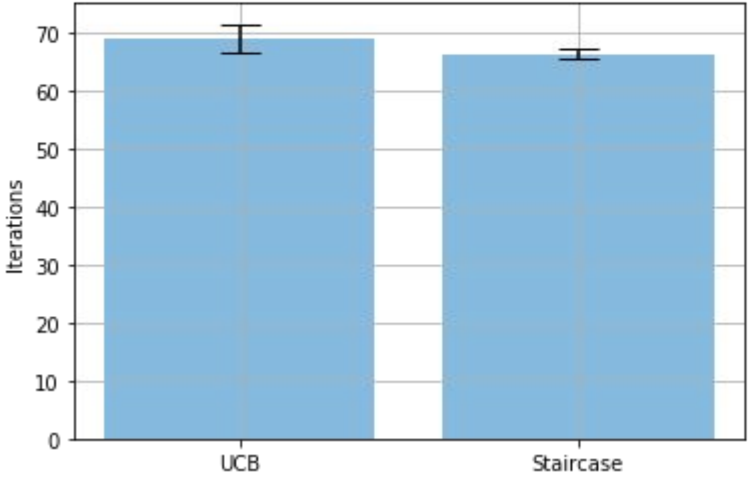


- Conservativeness: % of time subject does not notice the distortion when they should.
- UCB - Most robust staircase
  - No significant difference
- UCB - Least conservative staircase:
  - Significant difference
- UCB - In-Middle staircase:
  - No significant difference

⇒ UCB's conservativeness is in between in-middle staircase and most robust staircase.

⇒ H2 verified for least robust staircase.

Criterion 3 - Convergence speed



- UCB: 69 +- 11.5 iterations
- Staircase: 66 +- 4 iterations
- 1 trial : 6.5 +- 0.57s
- Staircase converges 1min35s faster than UCB

** 25% of staircases did not converge

⇒ Staircase converges faster, but the difference in time is negligible.

⇒ H3 rejected.

Conclusion

- Robustness: UCB is as robust as the most robust staircase
  - H1 ⇒ Verified for least conservative staircase
- Conservativeness: UCB is as conservative as the most robust staircase
  - H2 ⇒ Verified for least conservative staircase
- Convergence speed: Staircase converges in less iterations than UCB. UCB takes 1 min 35s more time.
  - H3 ⇒ Rejected
- Adaptivity: UCB is by definition adaptive. Staircase is not adaptive
  - H4 ⇒ Accepted

⇒ UCB is a better algorithm to use for our experiment.

Limits

- 25% of staircases did not converge.
  (95% UCB converged).
- Increase the maximum number of iterations (20) for staircase to have better convergence probability.
- Verify the adaptivity of UCB with subjects
  - Tests in ideal conditions show promising results

Future work

- EEG: take implicit feedbacks from the subject (Su-Kyoung 2017)
- Motor rehabilitation: help the subjects to consider the movement as their own and may accelerate recovery in the motor rehabilitation (Cameirao 2011)

References

- Konstantina Kilteni, Raphaela Groten, and Mel Slater. The sense of embodiment in virtual reality. *Presence Teleoperators amp Virtual Environments*, 21, 11 2012.
- Porssut, B. Herbelin, and R. Boulic. Reconciling being in-control vs. being helped for the execution of complex movements in vr. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pages 529–537, March 2019.
- Henrique Galvan Debarba, Ronan Boulic, Roy Salomon, Olaf Blanke, and Bruno Herbelin. Self-attribution of distorted reaching movements in immersive virtual reality. Computers amp; Graphics, 76:142–152, 2018
- Sidney Bovet, Henrique Galvan Debarba, Bruno Herbelin, Eray Molla, and Ronan Boulic. The critical role of self-contact for embodiment in virtual reality. IEEE Transactions on Visualization and Computer Graphics, 24(4):1428–1436, April 2018
- Eric Burns, Sharif Razzaque, Abigail Panter, Mary Whitton, Matthew McCallus, and Frederick Brooks, Jr. The hand is more easily fooled than the eye: Users are more sensitive to visual interpenetration than to visual-proprioceptive discrepancy. Presence, 15:1–15, 02 2006.
- Inaki Iturrate, Luis Montesano, and Javier Minguez. Robot reinforcement learning using eeg-based reward signals.
- Tian-jian Luo, Ya-chao Fan, Ji-tu Lv, and Changle Zhou. Deep reinforcement learning from error-related potentials via an eeg-based brain-computer interface. pages 697–701
- Timothy Meese. Using the standard staircase to measure the point of subjective equality: A guide based on computer simulations. Perception psychophysics , 57:267–81, 04 1995
- Su-Kyoung Kim, Elsa Kirchner, Arne Stefes, and Frank Kirchner. In- trinsic interactive reinforcement learning – using error-related potentials for real world human-robot interaction. *Scientific Reports*, 7, 12 2017.
- Monica Cameirao, Sergi Bermudez i Badia, Esther Duarte, and Paul Verschure. Virtual reality based rehabilitation speeds up functional recovery of the upper extremities after stroke: A randomized controlled pilot study in the acute phase of stroke using the rehabilitation gaming system. Restorative neurology and neuroscience, 29:287–98, 05 2011.

Criterion 4 - Adaptivity

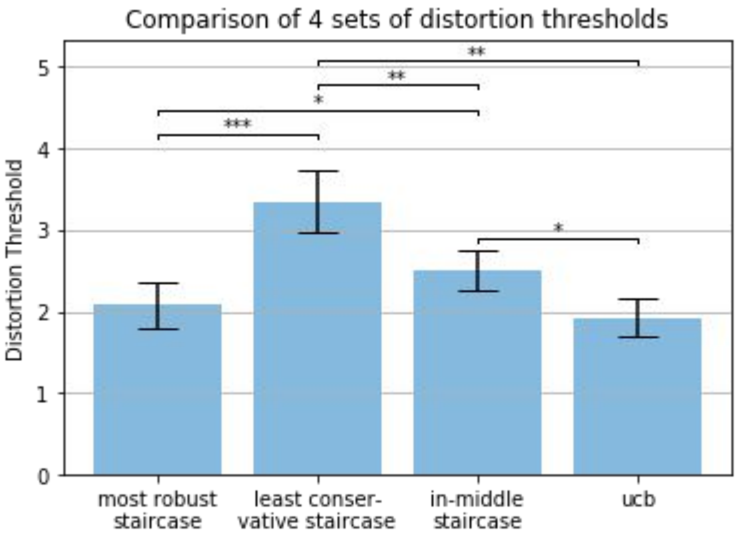
- No experience done with subjects
- UCB solves **non-stationary** multiarmed-bandit ⇒ Adaptive by definition
  - Tests in ideal conditions show promising results
- Staircase is not adaptive

Subjects’ comments

When do you notice the distortion? What factors causes you to notice it?

- Most of them notice the “jump” between the blue sphere to the green sphere
- Stability of the hand when following the green sphere
- Feel attracted / helped when leaving the blue sphere

Distortion Thresholds



- Distortion threshold: maximum magnitude of distortion without provoking BIE
- UCB - Most robust staircase
  - **No** significant difference
- UCB - Least conservative staircase:
  - **Significant** difference
- UCB - In-Middle staircase:
  - **Significant** difference

⇒ Significant difference between In-Middle and UCB, but not for Robustness (%) and Conservativeness(%)

UCB: Derivation of incremental update

Goal : Maximize expected reward

$$Q_*(a) = E[R_t | A_t = a]$$

Non-stationary : exponential, recency weighted average

$$Q_{n+1} = Q_n + \alpha\Big[R_n - Q_n\Big]$$
$$= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$$

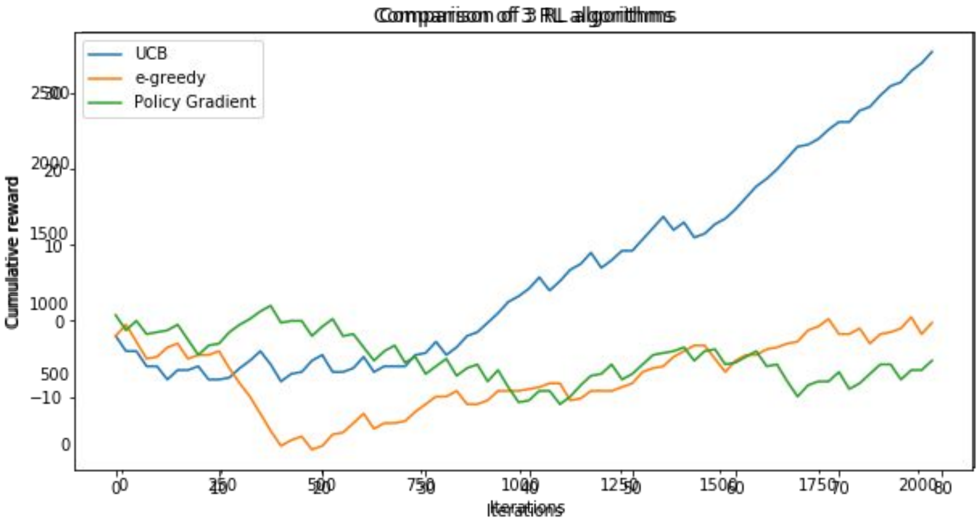
where α is a constant, *step-size parameter*, $0 < \alpha \leq 1$

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

$$\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big],
\end{aligned}$$

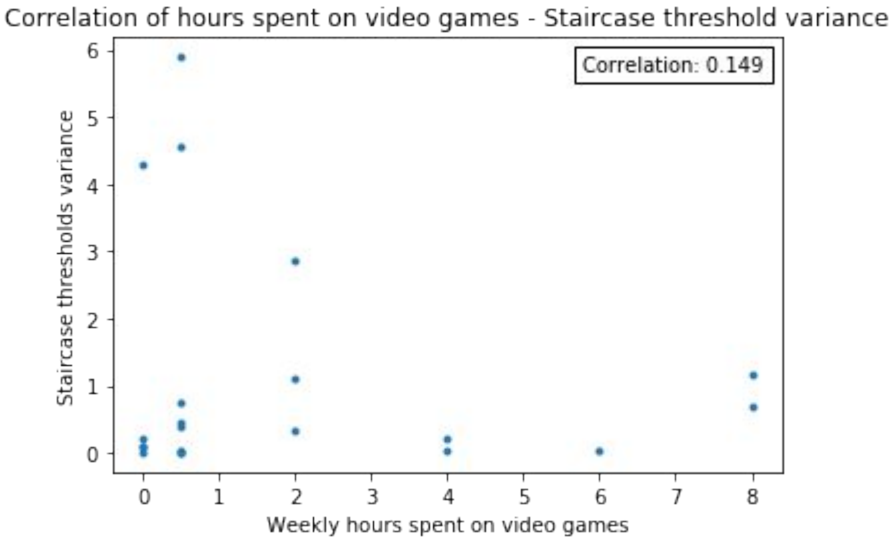
Performance of RL algorithms

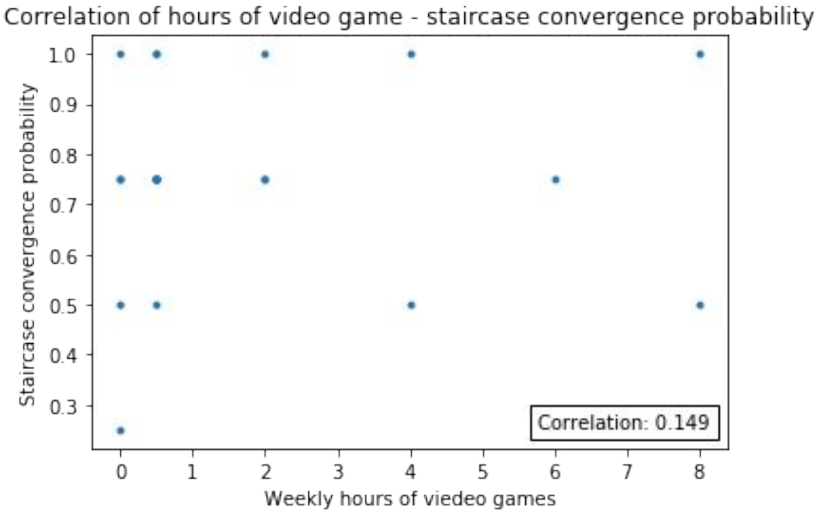
- Tests on 1000 trials
- Correct threshold:1.5
- Distortion values: [0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2]
- Converge when for 10 times the top action stays the same



	Iterations	Correctness
UCB + e-Greedy	45 +-11	~97%
e-Greedy	63 +- 23	~90%
Policy Gradient	~185+-54	20~40%
Staircase	80	N/A

Correlation with video gaming

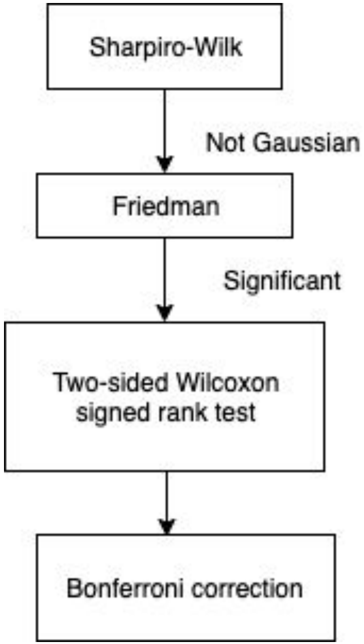




Why choose to explore RL

1. Original design: Use multiple parameters to build a complex model
    - Hand speed, EyeTracking, Hand orientation ⇒ Distortion value
2. Online adaptive model
    - Need an algorithm that would dynamically change the value of the distortion based on subject’s reaction
    - Point of Subjective Equivalence (fitting a curve) ⇒ Offline
    - Need to retrain staircase / PSE if the threshold varies over time
3. Rapidity of convergence:
    - Staircase: needs 80 and doesn’t necessarily achieve convergence
    - PSE: Convergence precision proportional to number of data points needed

Statistical Analysis



1. Shapiro-Wilk: Normality test
2. Friedman: non-parametric test for testing if a difference exists between several related samples
3. Wilcoxon: multiple comparison
4. Bonferroni: Correction for multiple comparison

Q-learning and SARSA

" Algorithm to learn a policy that will tell us how to interact with an environment under different circumstances in such a way to maximize rewards."

- Model-free: no transition table (i.e. prediction of what the next state will be)

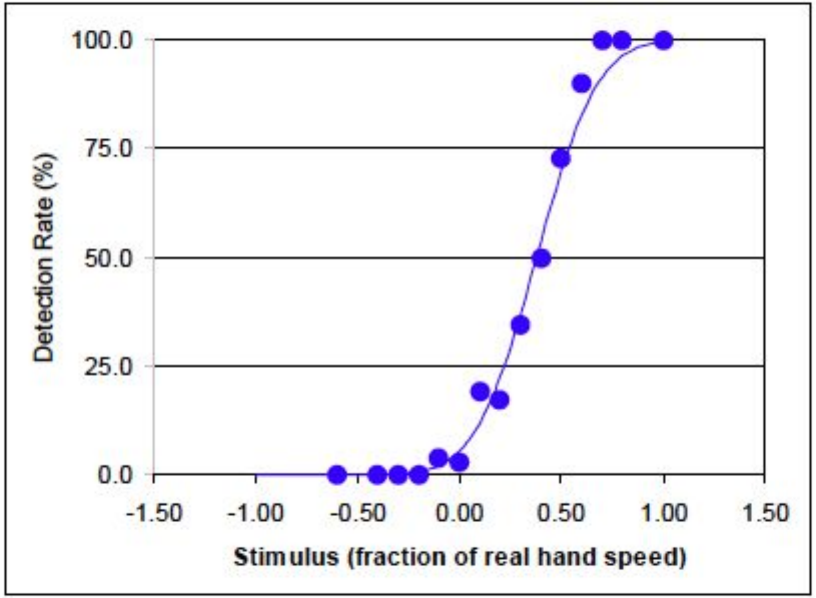
$$\text{NewEstimate} \leftarrow \text{OldEstimate} + \text{StepSize}\Big[\text{Target} - \text{OldEstimate}\Big]$$

Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Choose A from S using policy derived from Q (e.g., ϵ-greedy)
  Repeat (for each step of episode):
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ϵ-greedy)
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

"SARSA will approach convergence *allowing* for possible penalties from exploratory moves, whilst Q-learning will ignore them. That makes SARSA more conservative - taking less risk to trigger a large negative rewards towards the end" (https://stats.stackexchange.com/questions/326788/when-to-choose-sarsa-vs-q-learning)

2 staircase methods - Burns



- Use staircase to gather data
- Use a psychometric function fit to users' data points
- Distortion threshold: 50% detection