Aprendizado por Reforço

AULA - 1

Conceitos e Modelagem

Anteriormente...

- Otimizar: Encontrar a melhor solução para um problema
  - Maximizar ou minimizar uma função é encontrar os melhores valores para suas variáveis
- Aprendizado de máquina: Reconhecimento de padrões e fazer associações sem programação explícita
- Redes Neurais Artificiais: Conjuntos de neurônios paralelos e sequenciais cujas variáveis são alteradas durante o treinamento.
- **Imitation Learning**: Aprender comportamentos a partir de exemplos de experts
  - Problemas sequenciais precisam de representatividade nos dados, especialmente para correção de caminhos errados.

ROTEIRO

- O que é Aprendizado por Reforço
- Conceitos de Aprendizado por Reforço
  - Relação Agente e Ambiente
- Modelagem de problemas de Reforço
  - Identificando conceitos

O que é aprender por REFORÇO?

- Psicologia
  - Behaviorismo
- Adestração Animal

O que é aprender por REFORÇO?

- Construir associações através de estímulos positivos e negativos

- Aprender O QUE FAZER e QUANDO FAZER para maximizar os estímulos positivos

Mapeamento Estados e Ações

Estado/Situação	Melhor Ação
Ouvir “Senta!”	Sentar
Ouvir “Pega!”	Atacar Alvo
Ouvir “Dá a pata”	Erguer pata dianteira

Aprendizado Supervisionado

- Exemplos de comportamento correto
  - Função de perda proporcional aos erros

- Representatividade
  - Estabilização
  - Generalização

Aprendizado por Reforço

- Sinal de Recompensa
  - Função de perda proporcional ao sinal de recompensa
  - Granularidade
- Sequencial
  - Ações são tomadas em sequência
  - Recompensa Atrasada
- Interativo
  - Ações afetam as próximas situações
  - Tentativa e Erro
  - Experiência

“In interactive problems it is often impractical to obtain examples of desired behavior that are both correct and representative of all the situations in which the agent has to act.”

Richard S. Sutton and Andrew G. Barto - Reinforcement Learning: An Introduction - Second Edition, Pg. 2

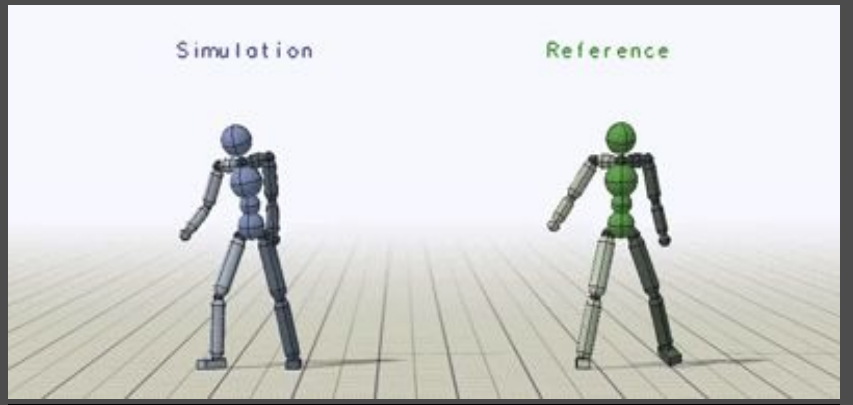
Outros Exemplos

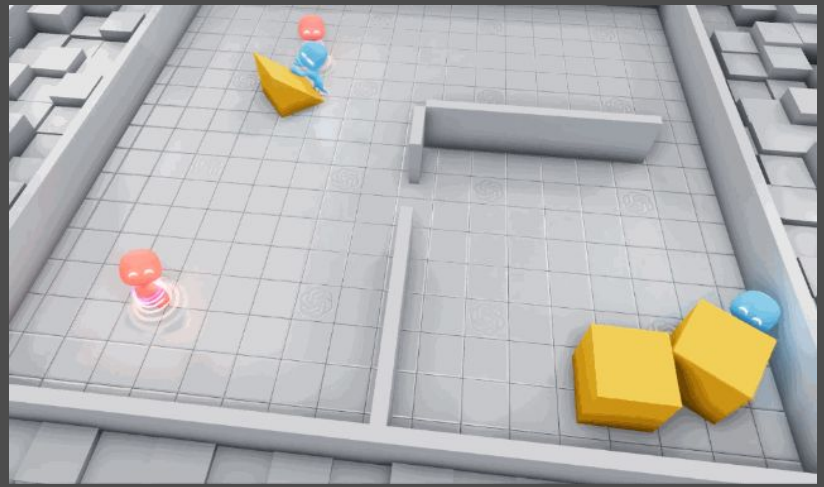
- Uma gazela bebê cambaleia e cai ao nascer, mas poucos minutos depois é capaz de correr a mais de 30 Km/h.

- Um controlador adaptativo ajusta os parâmetros de uma refinaria otimizando custos/qualidade.

- Um robô de limpeza decide entre adentrar em uma sala em busca de sujeira, ou retornar à estação de recarga.

Outros Exemplos





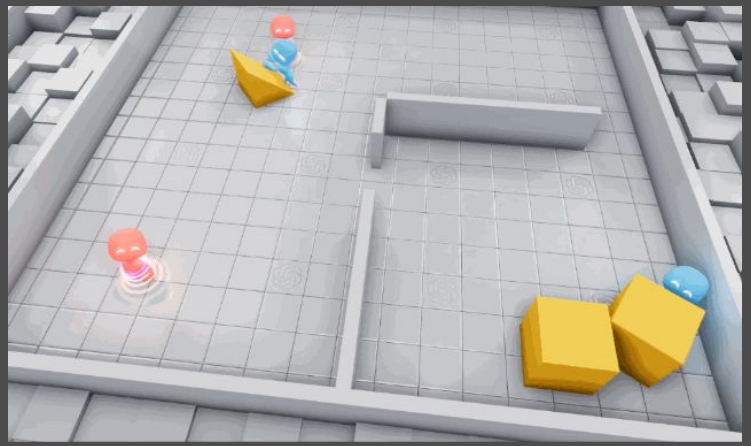
Relação Agente e Ambiente

Agente

- Atribuições:
  - Processar Observações
  - Agir
  - Aprender/Otimizar
- Subdivisões/Módulos
  - Política
  - Função de Valor
  - Modelo

Ambiente

- Elemento com o qual o agente irá interagir.
  - Engloba tudo que não é o Agente
  - Inclusive possíveis outros agentes



Observação/Estado

- Descrição da situação atual do ambiente
  - Pode ser parcial
  - Leitura de um sensor
  - Imagem
- Conjunto de todos os estados possíveis:
- Espaço de Estados

Observação deve permitir a inferência da recompensa

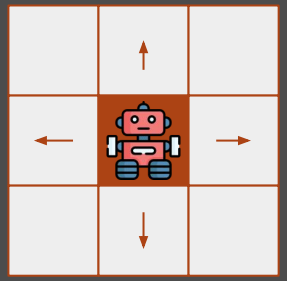
Espaço de Estados

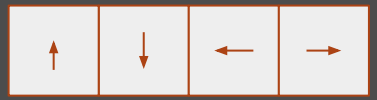
- Cada leitura (que assume valores diferentes) é uma dimensão no espaço
- 3 sensores (distâncias) = 3 dimensões
- Imagem 30x30x3 (RGB) = 2700 dimensões
  - O agente não verá TODAS as imagens possíveis do mundo

- Interessante olhar a dimensionalidade
- Noção de complexidade do problema

Ações

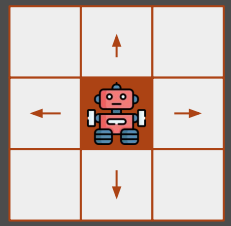
- Forma do Agente interagir com o Ambiente
- Conjunto de todas as ações possíveis:
- Espaço de Ações





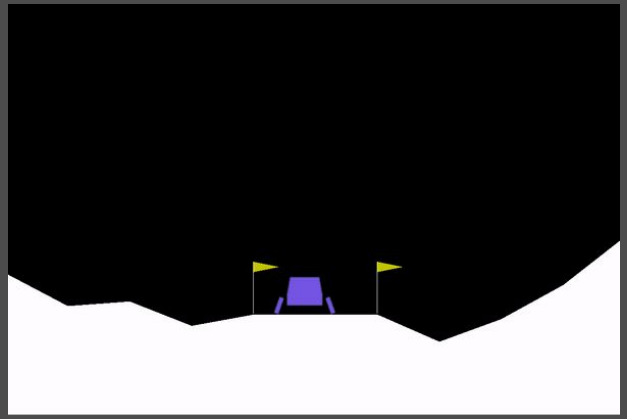
Espaço de Ações

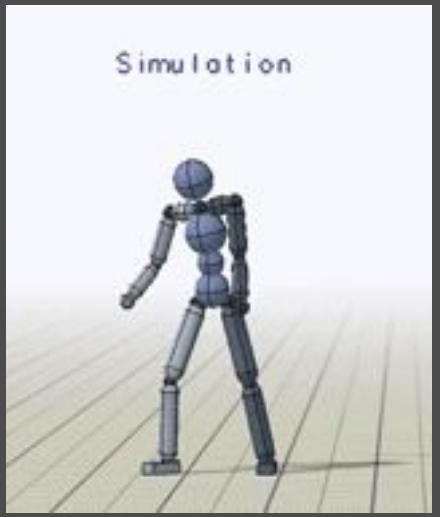
- Escolher uma dentre N ações: espaço unidimensional



- Escolher mais de uma ação simultânea: uma dimensão para cada ação que será escolhida

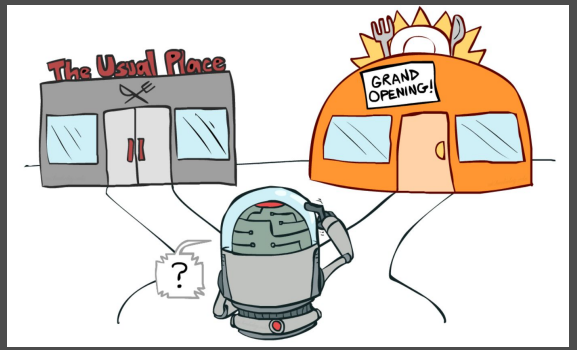
Qual o Espaço de Ações para estes problemas?



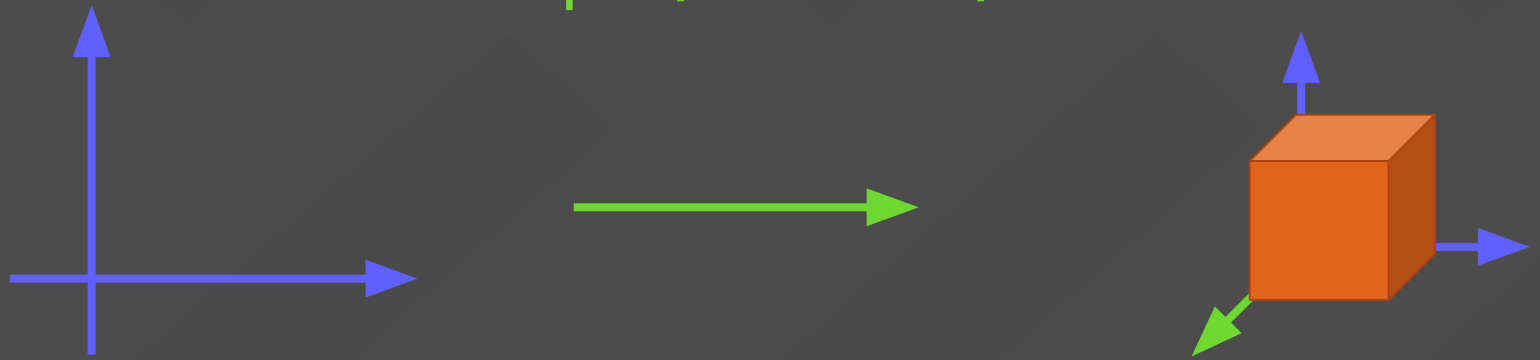


Exploration vs **Exploitation**

- **Exploration**: Encontrar mais informação sobre o ambiente, é preciso testar ações novas.
- **Exploitation**: Utilizar da informação adquirida previamente para maximizar a recompensa, é preciso maximizar a recompensa.
- Quanto mais opções de ações, mais difícil fica de explorar soluções.
- Quanto mais situações diferentes, também fica mais difícil de explorar o ambiente.



A complexidade do problema vai ser definida pela soma das dimensões do espaço de estados e do espaço de ações.



A questão não é apenas quantas opções eu tenho, mas sim em quantas situações eu terei que escolher dentre essas opções

Sinal de Recompensa

- Dizer o quão bom é o estado do ambiente.
- Imediatamente após a entrada no estado.
- "Atrelado" à última ação tomada.
- Diz o quão boa foi aquela ação específica?
  - Não necessariamente

s, a, r, s'

Transição

Função de Recompensa

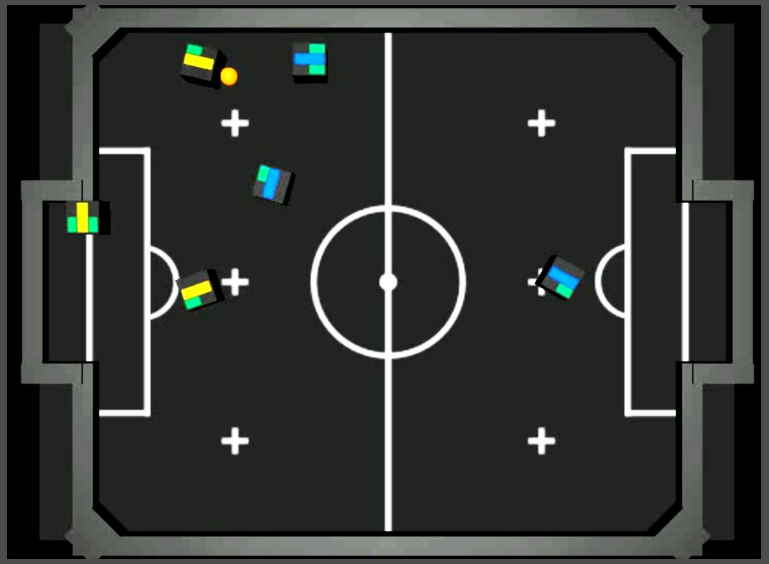
- Definida pelo programador
- Atrelada ao Ambiente
- Guia o processo de Aprendizado
- Esparsa, Contínua, ou qualquer meio termo
  - Quanto mais esparsa, menos *feedback* e mais liberdade
  - Quanto mais contínua, mais *feedback* e menos liberdade
  - Quanto mais *feedback*, mais rápido o aprendizado

Função de Recompensa

Qual seria uma boa função de recompensa para um jogo de xadrez?

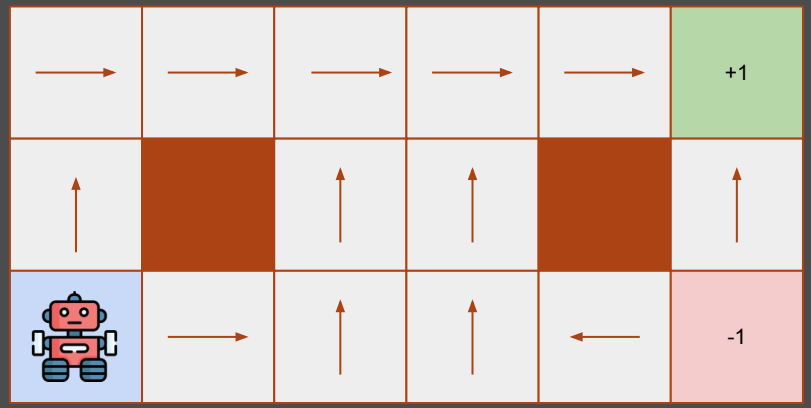
Função de Recompensa

Qual seria uma boa função de recompensa para um jogo de futebol de robôs?



Política

- Comportamento do Agente
- COMO ele mapeia estados a ações
- Representa uma solução para o problema
- O que muda durante o treino

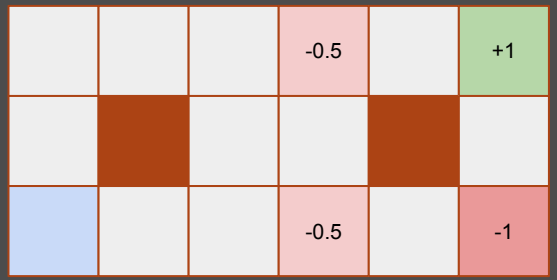


Política

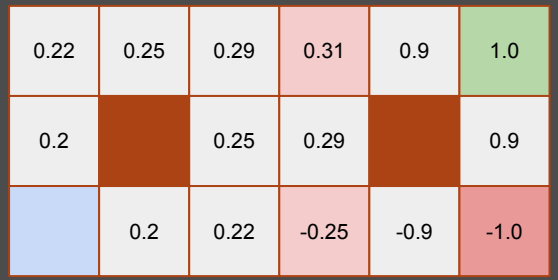
- Determina onde o Agente se encontra no processo de busca
  - Objetivo é encontrar a melhor política de decisão
  - Espaço de Busca ~ Espaço de Políticas
  - Espaço de Ações + Espaço de Estados + 1 (dimensões)
- Pode ser determinística ou estocástica
- Representada por uma simples tabela ou até redes neurais profundas

Função de Valor (opcional)

- Busca estimar o valor "real" de um estado (ou par estado-ação), considerando possíveis recompensas futuras.
- Depende da política
  - Também aprende com experiência



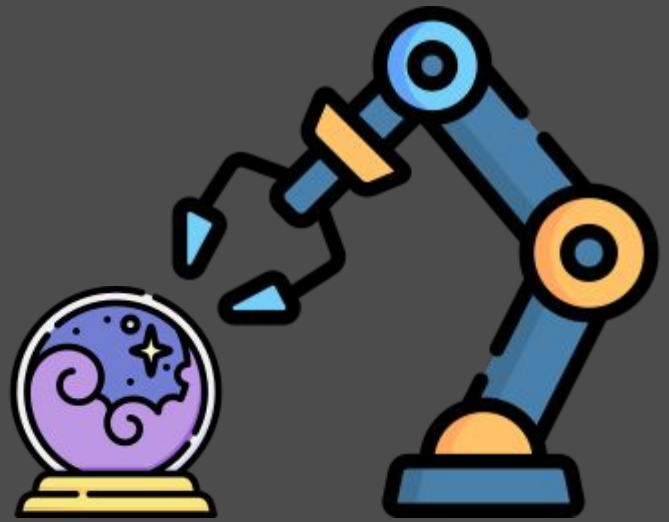
Recompensas



Valores

Modelo (opcional)

- Tenta predizer o próximo estado
- “Modelo” do ambiente
- Usados para planejamento
- Auxilia a política
- *Model-based*
- *Model-free*



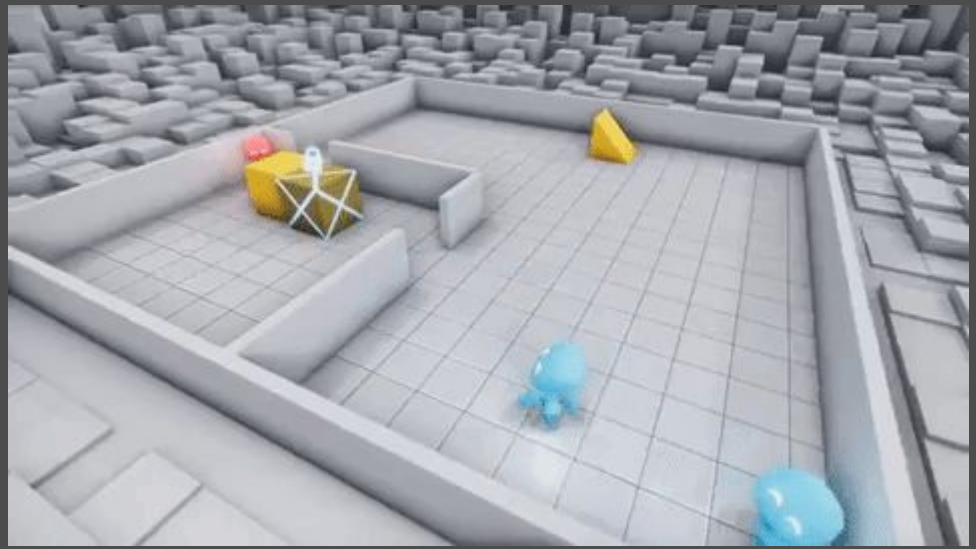
Algumas Terminologias

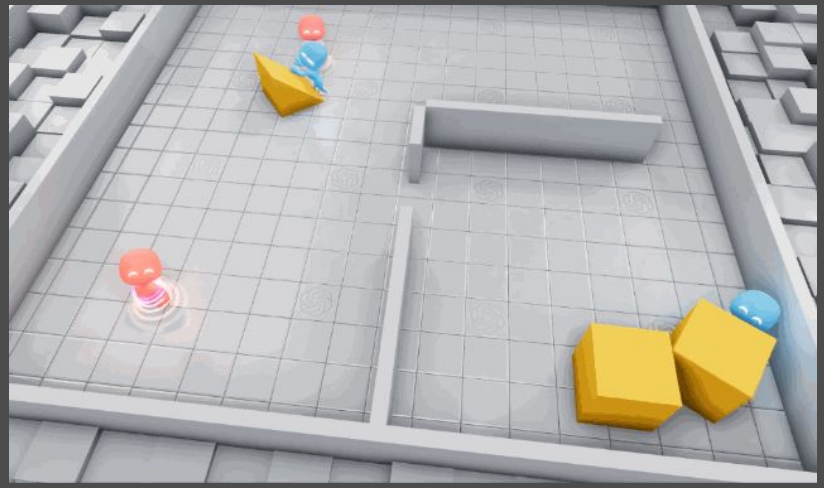
- *Policy Gradient/Improvement/Optimization*
  - Otimiza diretamente a política de decisão (probabilidade de ações)
- *Value-Based*
  - Otimiza apenas a função de valor, política segue a função de valor
- *Model-Based*
  - Otimiza um modelo do ambiente para planejar decisões
  - Geralmente usado em conjunto com outra abordagem

Modelar um problema de Reforço

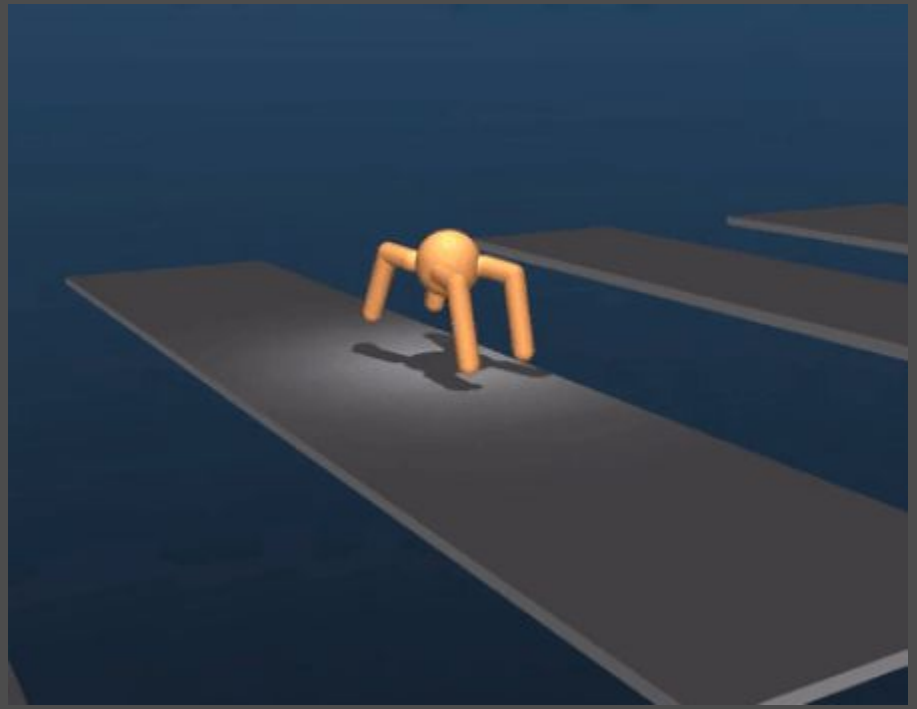
- O que é o meu Agente?
- O que é o meu Ambiente?
- Quais são as Ações possíveis?
  - Como eu represento meu espaço de ações?
- Quais são as Observações possíveis?
  - Como eu represento meu espaço de estados?
- Qual é a minha função de Recompensa?
  - Qual o objetivo?

Modelagens:

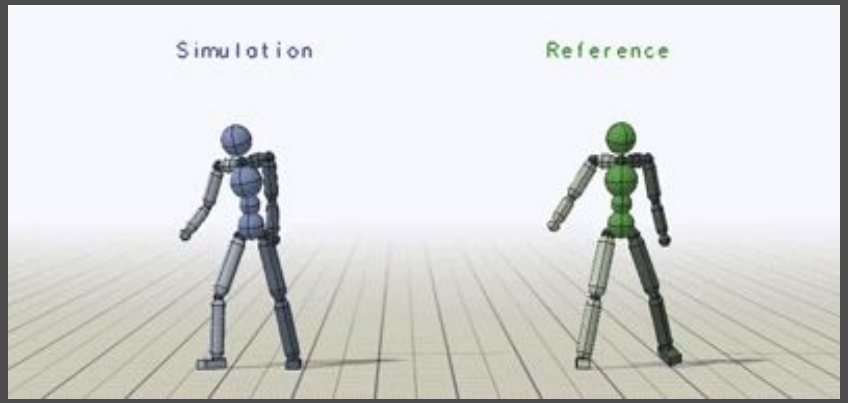


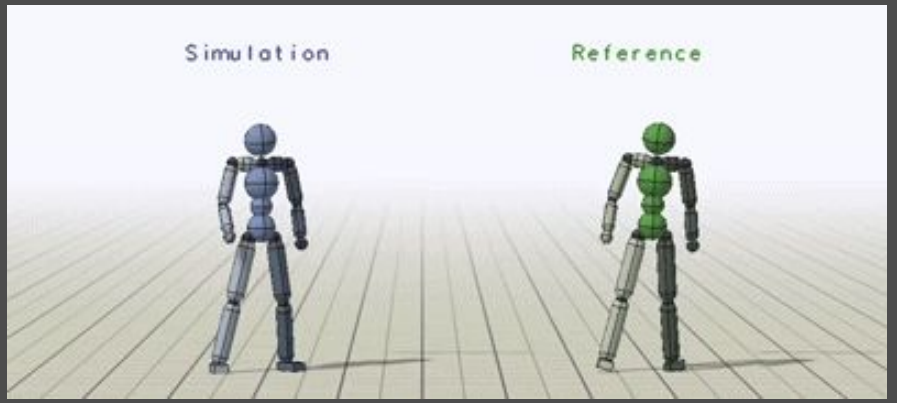


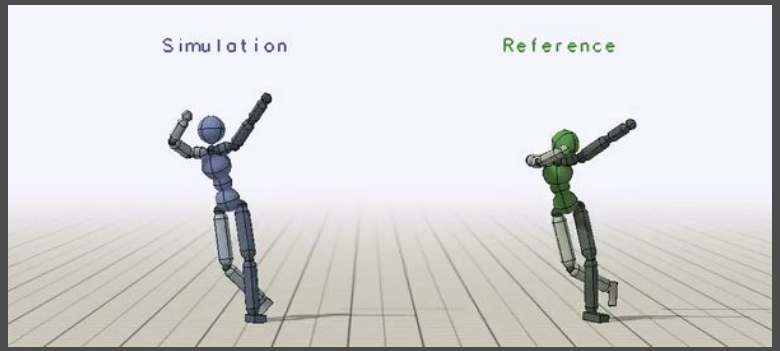
Modelagens:

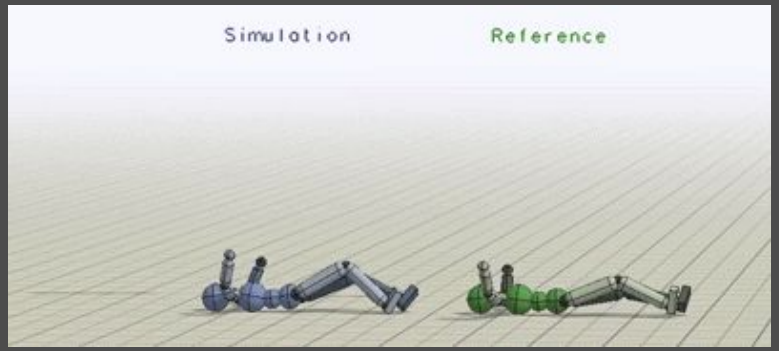


Modelagens:


Simulation
Reference


Simulation
Reference


Simulation
Reference


Simulation
Reference

É só...
Por enquanto