Clever Machines Learn How to Be Curious

Kubilay Agi

UID: 304784519

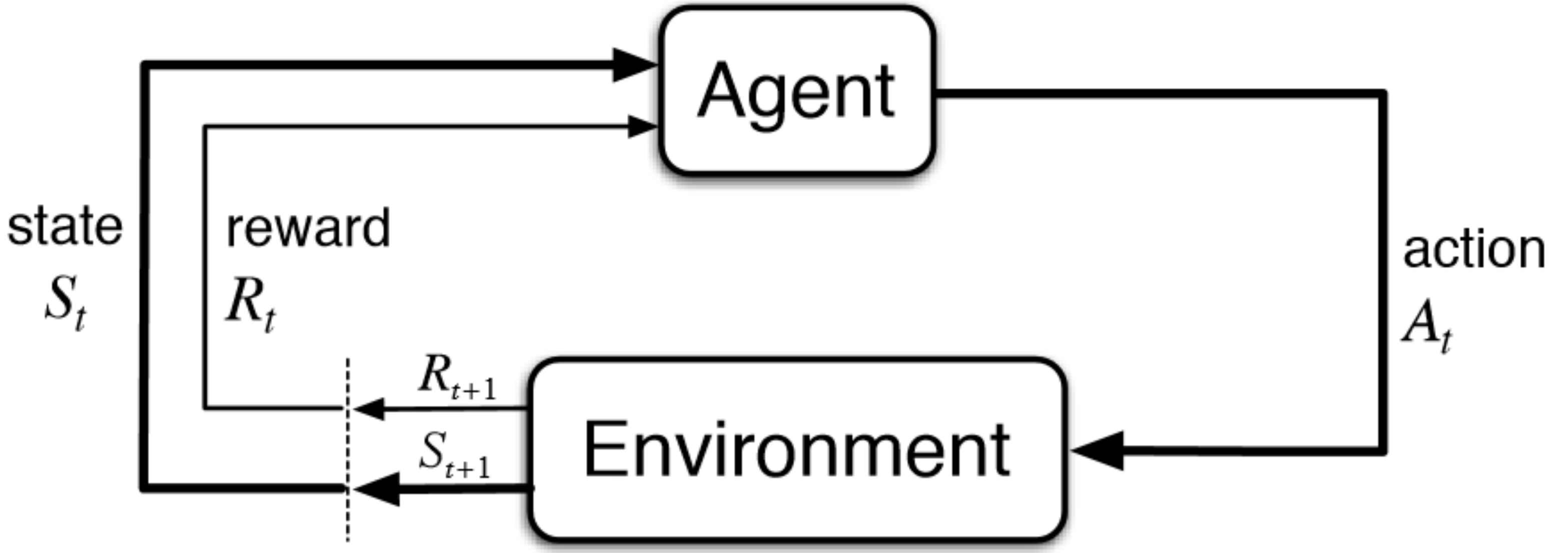
Curiosity and Reward Systems

- Intrinsic Motivation
  - Internal desire to learn
  - Ex: learning how to play a new video game

- Extrinsic Motivation
  - Point system
  - Ex: Salary

Point Systems in Machine Learning

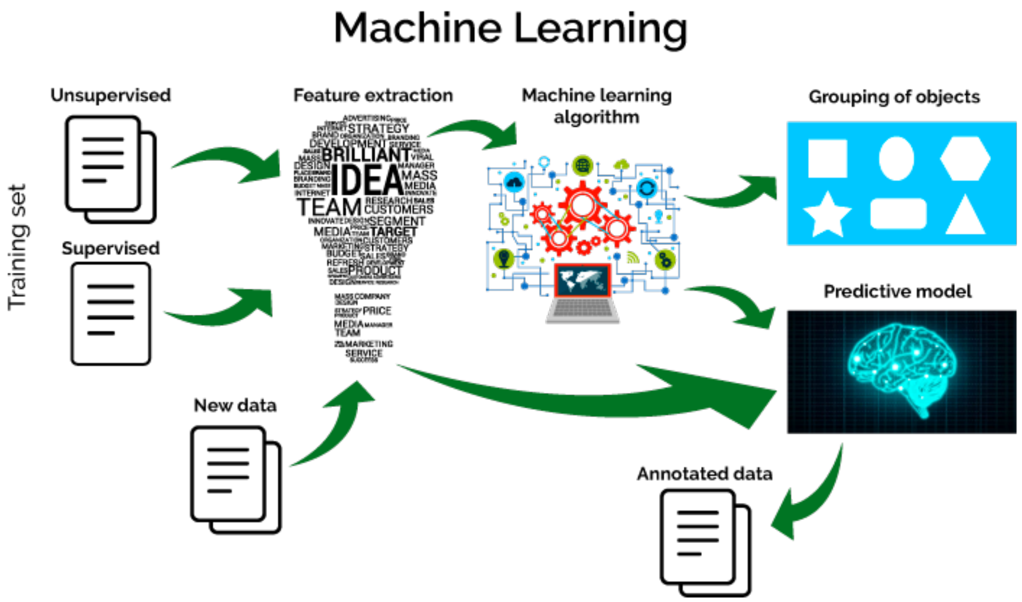
- Reward good behavior
- "Punish" bad behavior
- Repeat



Issues with this Point System

- Doesn’t allow for any grey-area for the machine
  - Everything must be hard coded as a reward or punishment
    - Not portable
    - Machine will get lost
  - Counterproductive for the purpose of machine learning

- Requires too much repetition
  - Ex: Autonomous cars don’t get second chances

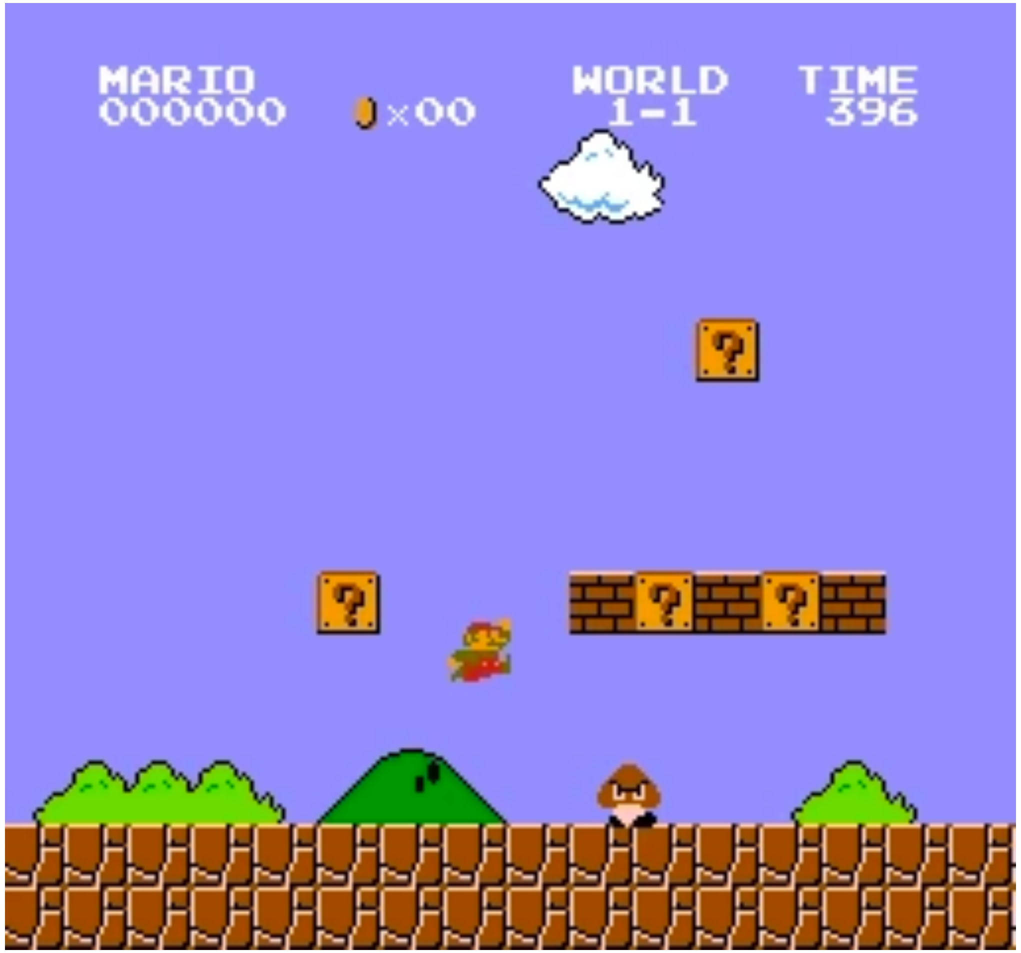


Intrinsic Curiosity

- Machine explores its environment - greedy

- Psychological study says infants prefer toys that are the most surprising

Applications



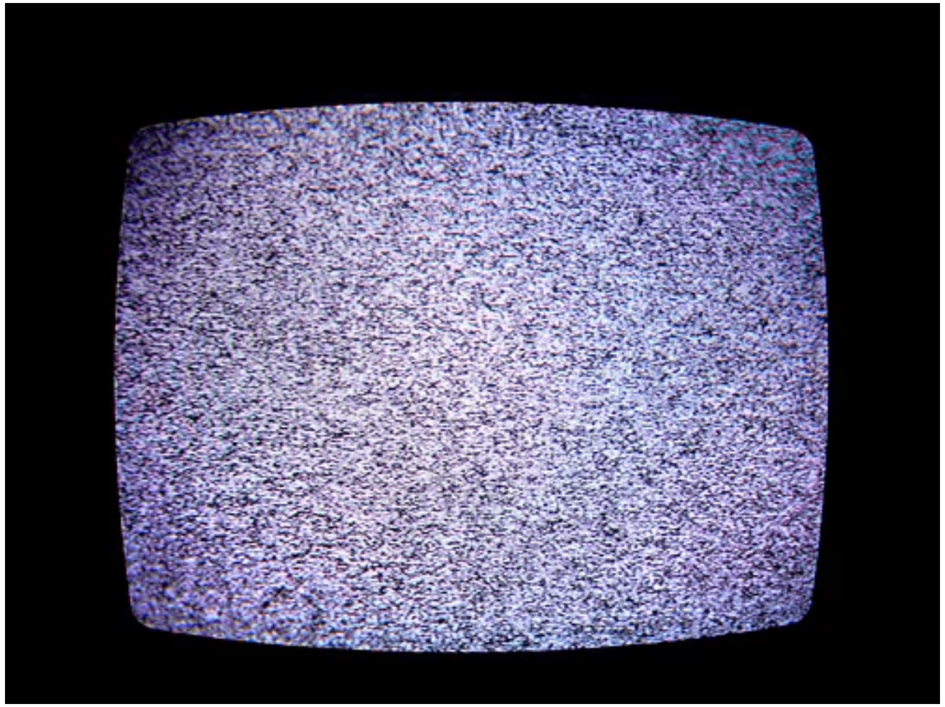
Intrinsic Curiosity in Machine Learning

- Pulkit Agrawal and Deepak Pathak based their machine learning system on this intrinsic curiosity and surprise driven learning

- In context of Super Mario Bros:
  - Develops mathematical representation of game
  - Predicts what game should look like in a few frames
  - Intrinsic reward signal based on how wrong the model was
  - “The higher the error rate — that is, the more surprised it is — the higher the value of its intrinsic reward function.”
    - Drawn toward unexplored states

Intrinsic Curiosity Models

- Nothing is predictable in the real world
  - Ex: leaves blowing across the ground – image constantly changing
  - Can’t be modeled as easily as pixels in a game



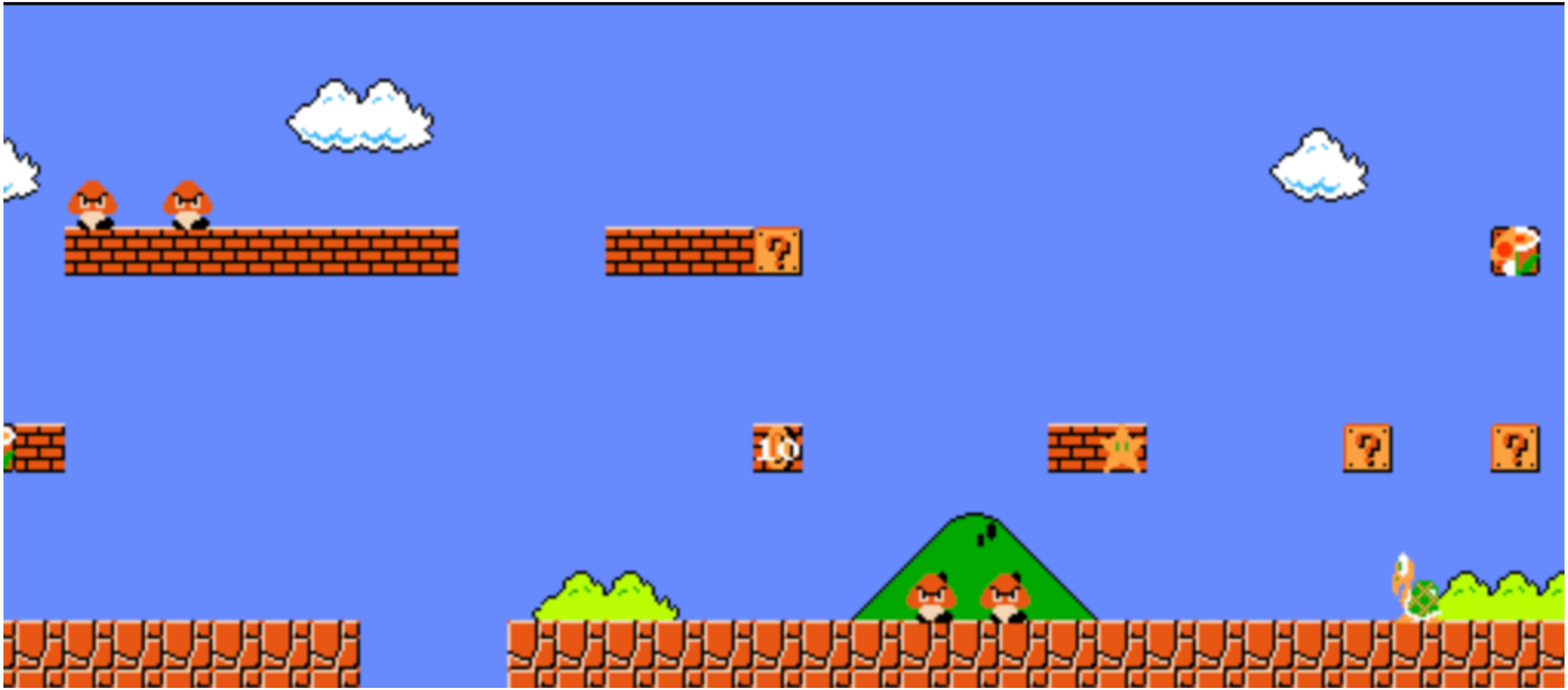
- Solution: Be curious, but not too curious.
  - Still searching for right balance

Improvements

- Abstraction:
  - The technology translates from individual pixels to general features

- Best of both worlds approach:
  - Combine extrinsic and intrinsic motivation styles to better guide the machine

AI’s Arch-Nemesis



Race Conditions

- Difficult to emulate human curiosity in machines

- Why? Psychological study does not have a formal definition for curiosity, or even a reason why we are curious in the first place.

- Offers possibilities of collaboration

A Simple Conclusion

- Better, but not perfect (yet)
- There is a time and place for everything
- A long way still to go

Bibliography

- Pavlus, John. *Clever Machines Learn How to Be Curious*. 19 Sept 2017. Quanta Magazine. https://www.quantamagazine.org/clever-machines-learn-how-to-be-curious-20170919/

- Barto, Andrew. *Intrinsically Motivated Learning of Hierarchical Collections of Skills.* http://www.lira.dist.unige.it/teaching/SINA_08-09/SINA_PREV/library/ICDL2004/pdfs/24.pdf

- Baldassarre, Gianluca. *Intrinsically Motivated Learning Systems: An Overview.* https://link.springer.com/chapter/10.1007/978-3-642-32375-1_1#Sec3