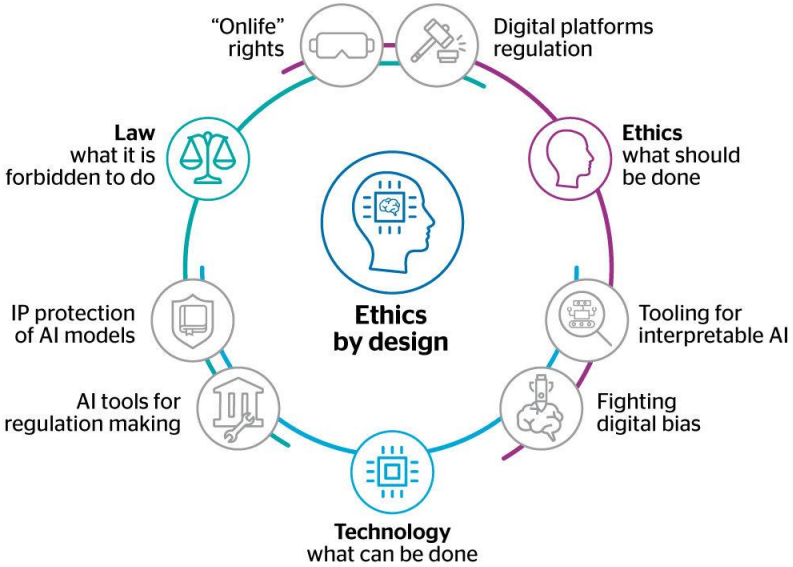
Ethics in Reinforcement Learning

Ethics

" Ethics is a branch of philosophy that involves systemizing, defending and recommending concepts of right and wrong behaviour"

- It has a subfield 'ethics of technology' which tries to answer if it is always, never, or contextually right or wrong to invent and implement a technological innovation. For example: computer viruses, nuclear weapons, environmental stability.

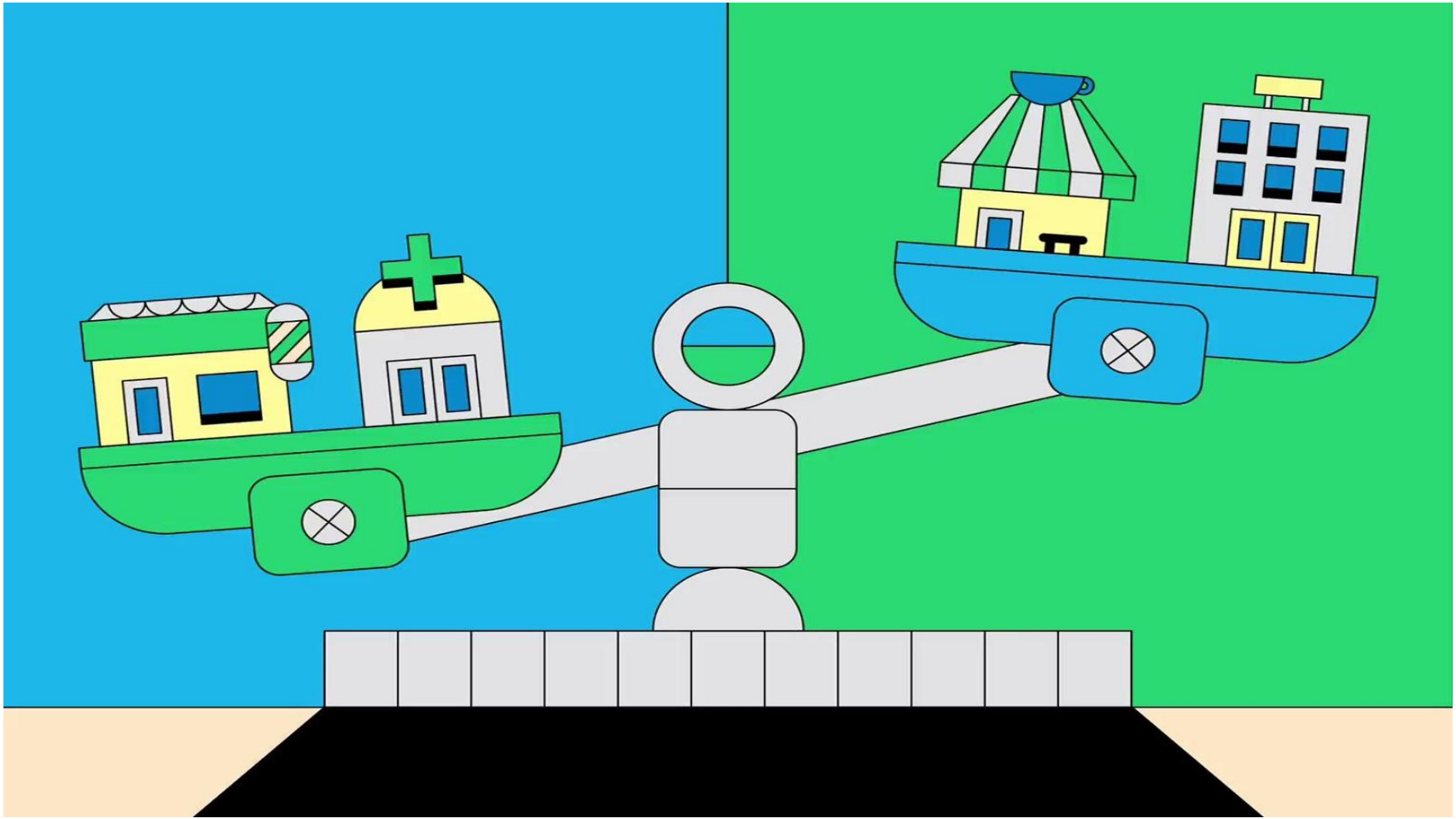


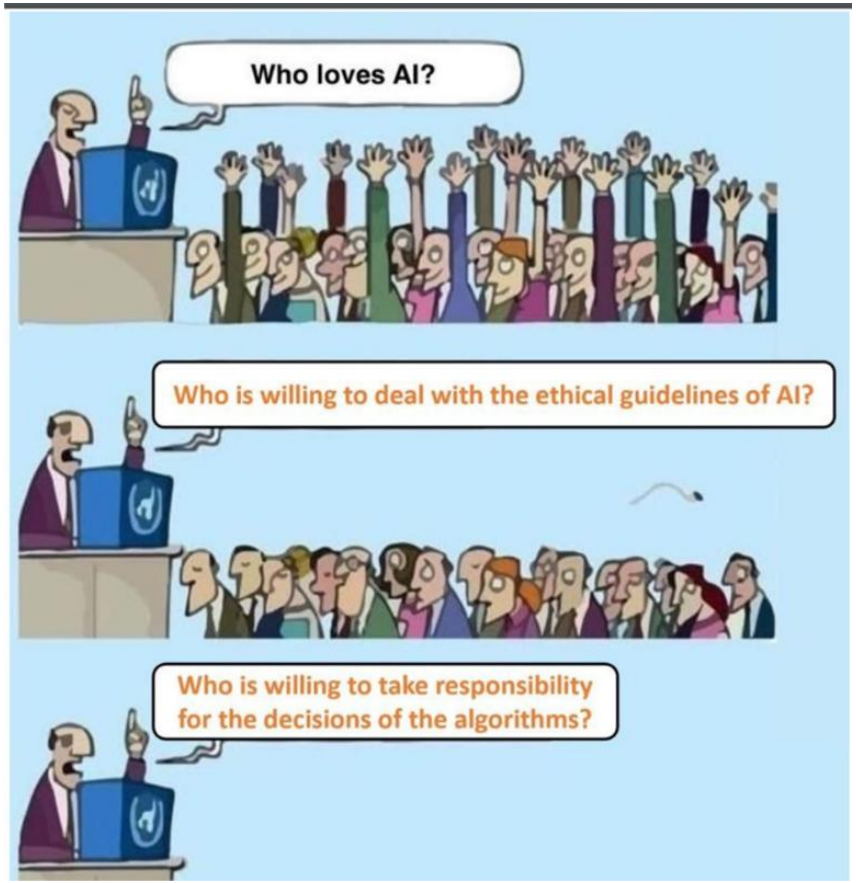
Ethics in AI

It is a part of 'ethics of technology' that is specific to robots, and other artificially intelligent beings.

It can be divided into:

- Roboethics: Deals with moral behaviour of human beings as they design, construct, use and treat artificially intelligent beings.
- Machine Ethics: Deals with the behaviour of Artificial Moral Agents (AMA).

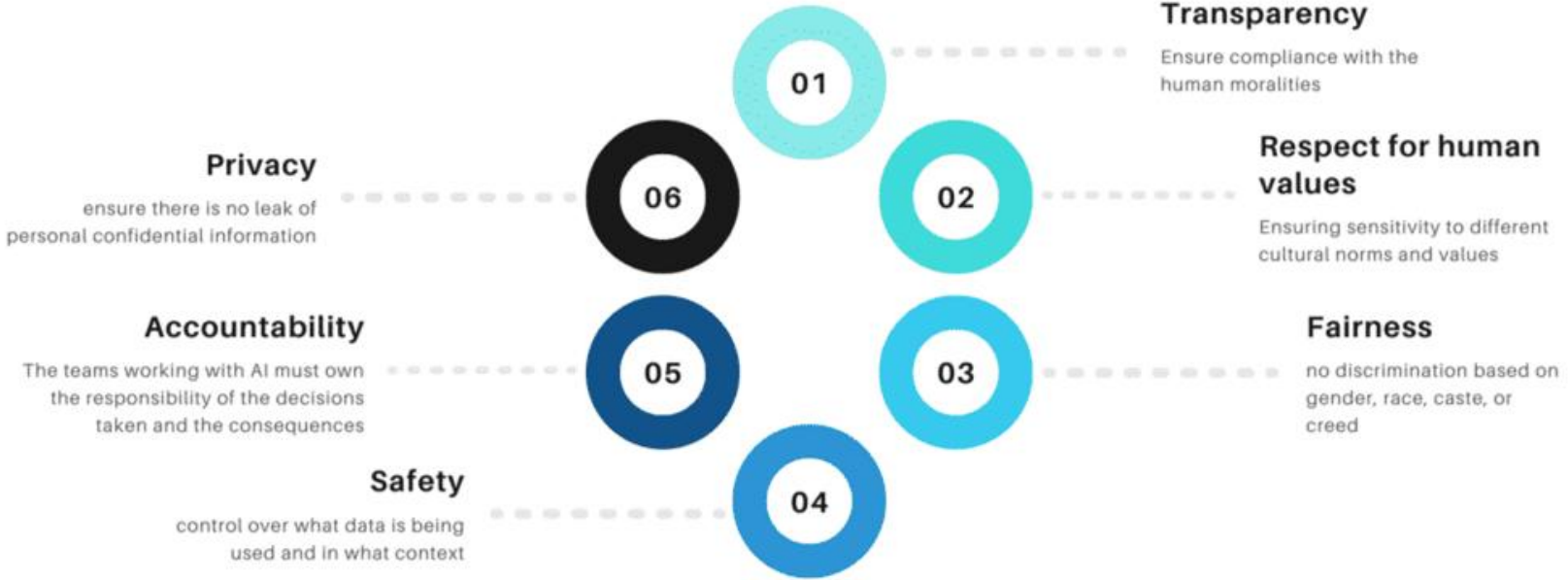



Who loves AI?
Who is willing to deal with the ethical guidelines of AI?
Who is willing to take responsibility for the decisions of the algorithms?

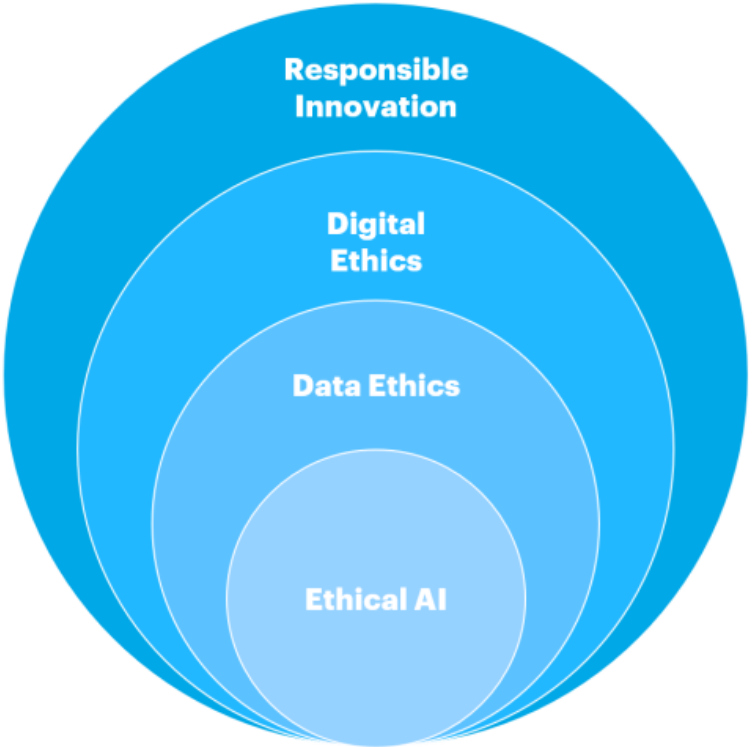

We should urgently develop an AI-Strategy!
Sure, but first we need to develop a Data-Strategy.

Ethics in AI

AI lacks awareness of a “moral compass” or empathy to judge what is right and what is wrong



Ethics in AI



Responsible Innovation means building trust in the future through collective stewardship of science, product, and service innovation in the present.

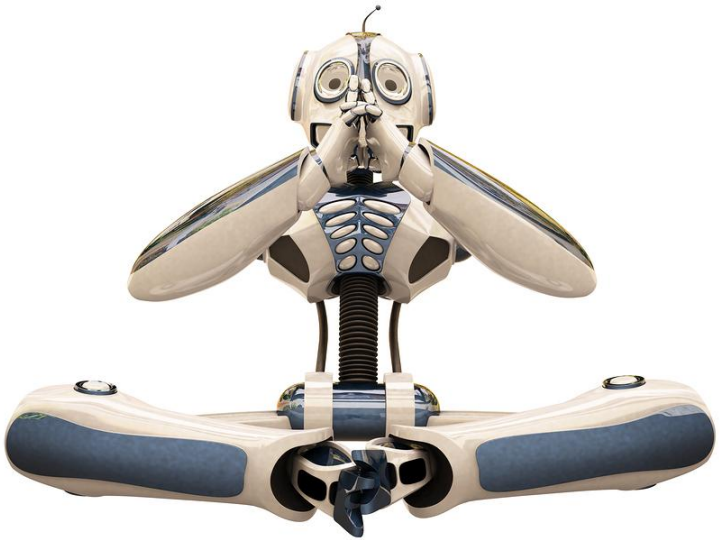
Digital Ethics is exhibited when systems of systems reinforce the work that is performed in service to responsible innovation to satisfy values, in accordance with principles, and in support of governance.

Data Ethics seeks to cure automated decision-support systems from the negative consequences of data quality issues with a systemic approach to integrity and provenance throughout the data supply chain.

Ethical AI is a discipline concerned with the fairness, bias, and efficacy of decision-support systems. AI ethicists can often treat symptoms of injustice that can emerge from automated systems.

Roboethics

- Coined by Gianmarco Veruggio in 2002, it focuses upon how artificially intelligent beings may be used to benefit or harm humans.
- It deals with so called Robot Rights, which are the moral obligations of society towards its machines, similar to human rights or animal rights.



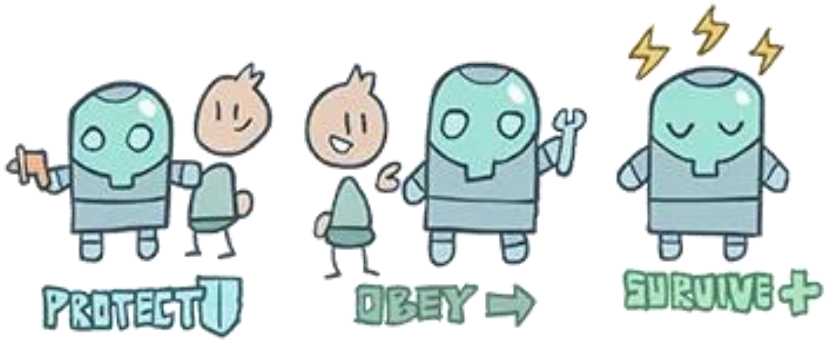
Machine Ethics

It focuses on how a machine learns new things – the algorithms involved (discussed later), can it learn on its own, can it become superior to humans etc.

THE LAWS OF ROBOTICS

The Three Laws of Robotics

1. A robot must not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.



The Zeroth Law

As the low-numbered law supersede the higher numbered laws, hence it gets its name.

It states that -

- A robot must not harm humanity, or, by inaction, allow humanity to come to harm.
- How does it decide in practice, as human beings are concrete objects but humanity is an abstraction – injury to it cannot be judged or estimated.
- A robot may not harm a human being, unless he finds a way to prove that ultimately the harm done would benefit humanity in general.

The Fourth Law

It states that -

- What happens if a human-like robot is made in future and the robot starts destruction in order to save this one human-like life.
- Hence the fourth law states that a robot must identify itself as a robot in all cases or in other words, a robot must know it is a robot.

THE NEED FOR AI ETHICS

The Need for AI Ethics

If AIs were to have no effective intrusion into our lives then whether they could be ethical would be of interest to only those who engage in thought experiments.

But AIs do intrude in our lives !

The kinds of intrusion are broadly of two sorts :

- Intrusions in which we have no choice but to interact with.
- Intrusions which cause and are a result of significant changes to culture and human interaction in particular.

The Need for AI Ethics

- An example of the first kind is : Self Driving Car.

- While ATMs come under the second category as they influence us by forcing to take currency of specified denominations

Turing Says …

"The original question, 'Can computers think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs."

AI Biases

The four most common types of BIAS seen in AI:

Reporting Bias: This type of AI bias arises when the frequency of events in the training dataset doesn't accurately reflect reality.

E.g : Take an example of a customer fraud detection tool that underperformed in a remote geographic region, marking all customers living in the area with a falsely high fraud score.

Selection Bias: This type of AI bias occurs if training data is either unrepresentative or is selected without proper randomization.

E.g., An example of selection bias is well illustrated by the research conducted by Joy Buolamwini, Timnit Gebru, and Deborah Raji, where they looked at three commercial image recognition products. The tools were to classify 1,270 images of parliament members from European and African countries. The study found that all three tools performed better on male than female faces and showed more substantial bias against darker-skin females, failing on over one in three women of color — all due to the lack of diversity in training data.

AI Biases

Group Attribution Bias: Group attribution bias takes place when data teams extrapolate what is true of individuals to entire groups the individual is or is not part of.

E.g., This type of AI bias can be found in admission and recruiting tools that may favor the candidates who graduated from certain schools and show prejudice against those who didn't.

Implicit Bias: This type of AI bias occurs when AI assumptions are made based on personal experience that doesn't necessarily apply more generally.

E.g., if data scientists have picked up on cultural cues about women being housekeepers, they might struggle to connect women to influential roles in business despite their conscious belief in gender equality — an example echoing the story of Google Images' gender bias.

Types of machine learning bias

- Algorithm bias : This occurs when there's a problem within the algorithm that performs the calculations that power the machine learning computations.

- Sample bias: This happens when there's a problem with the data used to train the machine learning model. In this type of bias, the data used is either not large enough or representative enough to teach the system.

    - For example, using training data that features only female teachers will train the system to conclude that all teachers are female.

Types of machine learning bias

- Prejudice bias : In this case, the data used to train the system reflects existing prejudices, stereotypes and/or faulty societal assumptions, thereby introducing those same real-world biases into the machine learning itself.
  - For example, using data about medical professionals that includes only female nurses and male doctors would thereby perpetuate a real-world gender stereotype about healthcare workers in the computer system.

- Measurement bias : As the name suggests, this bias arises due to underlying problems with the accuracy of the data and how it was measured or assessed.
  - a system being trained to precisely assess weight will be biased if the weights contained in the training data were consistently rounded up.

Types of machine learning bias

Exclusion bias : This happens when an important data point is left out of the data being used --something that can happen if the modelers don't recognize the data point as consequential.

How to prevent bias

1. Select training data that is appropriately representative and large enough to counteract common types of machine learning bias, such as sample bias and prejudice bias.
2. Test and validate to ensure the results of machine learning systems don't reflect bias due to algorithms or the data sets.
3. Monitor machine learning systems as they perform their tasks to ensure biases don't creep in over time as the systems continue to learn as they work.
4. Use additional resources, such as Google's What-if Tool or IBM's AI Fairness 360 Open Source Toolkit, to examine and inspect models.

MORAL EXPECTATIONS FROM AI

Qualities in AI

Qualities that we usually consider while making machines

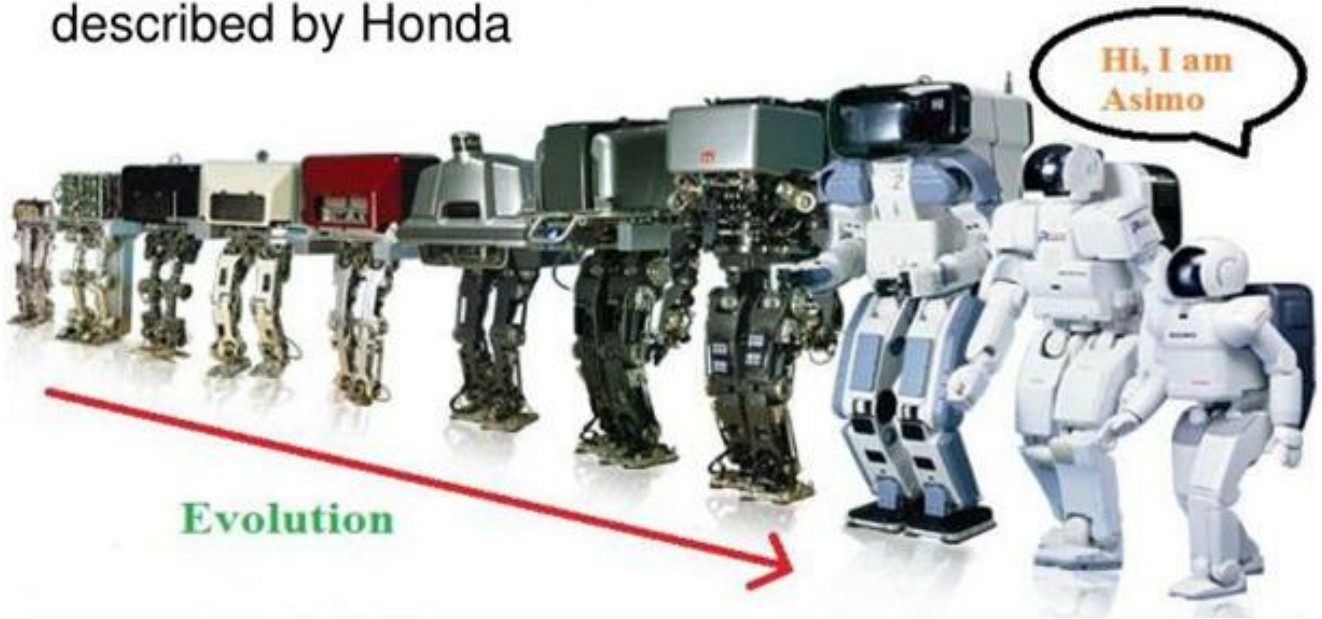
- Powerful (Speed, Accuracy etc.)
- Scalability – how an algorithm scales up on larger parallel systems

Qualities that we need to consider

- Transparency
- Auditability
- Incorruptibility
- Responsibility
- Predictability

Asimo

- Here is Asimo, the world's most humanoid robot as described by Honda



Asimo

- ASIMO has the ability to recognize moving objects, postures, gestures, its surrounding environment, sounds and faces, which enables it to interact with humans.
- The robot can detect the movements of multiple objects by using visual information captured by two camera "eyes" in its head and also determine distance and direction.
- The robot interprets voice commands and human gestures, enabling it to recognize when a handshake is offered or when a person waves or points, and then respond accordingly. ASIMO's ability to distinguish between voices and other sounds allows it to identify its companions.
- ASIMO is able to respond to its name and recognizes sounds associated with a falling object or collision. This allows the robot to face a person when spoken to or look towards a sound.
- ASIMO responds to questions by nodding or providing a verbal answer in different languages and can recognize approximately 10 different faces and address them by name.[15]

Some Ethical Issues

- Reporting of results
- Interpretability of algorithm behaviour
- Discrimination and bias learned from human data
- The possibility of Artificial General Intelligence

Reporting of results

- Statistical methodological issues
- Failure to report negative results.
- Cherry-picking easy tasks that look impressive.
- Failure to investigate performance properly.
- **Overall: the AI Hype problem!**

Requirements of Trustworthy AI

1 Human agency and oversight
Including fundamental rights, human agency and human oversight.

2 Technical robustness and safety
Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility.

3 Privacy and data governance
Including respect for privacy, quality and integrity of data, and access to data.

4 Transparency
Including traceability, explainability and communication.

5 Diversity, non-discrimination and fairness
Including the avoidance of unfair bias, accessibility & universal design, & stakeholder participation

6 Societal and environmental wellbeing
Including sustainability & environmental friendliness, social impact, society & democracy

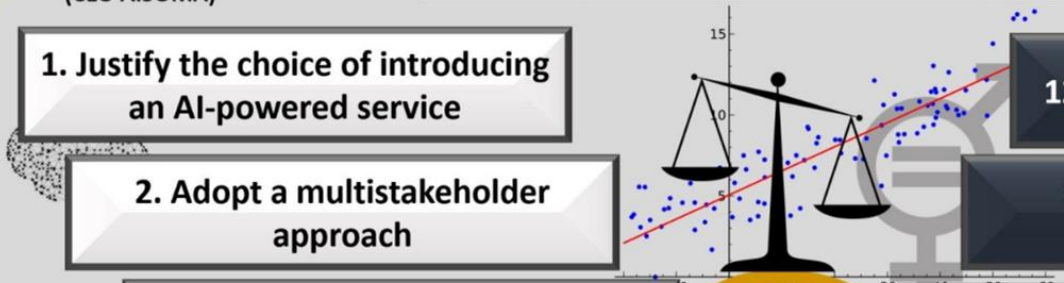
7 Accountability
Including auditability, minimisation & reporting of negative impact, trade-offs & redress

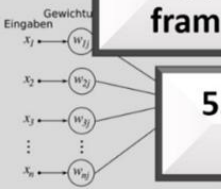
To be continuously evaluated and addressed throughout the AI system's Life Cycle

Murat Durmus
(CEO AISOMA)

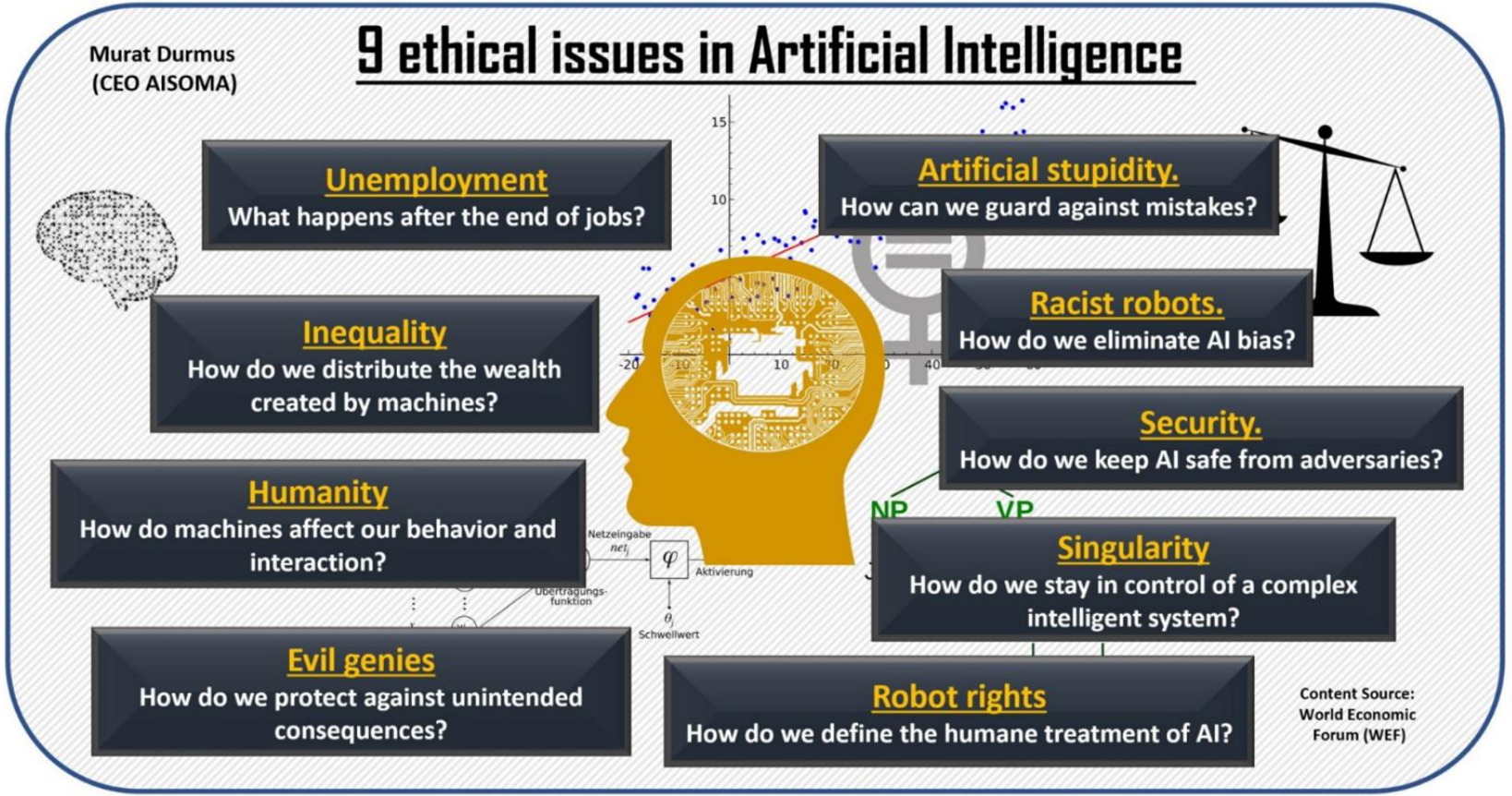
12 steps to put AI-Ethics into practice

1. Justify the choice of introducing an AI-powered service
2. Adopt a multistakeholder approach
3. Consider relevant regulations and build on existing best practices
4. Apply risks/benefits assessment frameworks across the lifecycle
5. Adopt a user-centric and use case-based approach
6. Clearly lay out a risk prioritization scheme
7. Define performance metrics
8. Define operational roles
9. Specify data requirements and flows
10. Specify lines of accountability
11. Support a culture of experimentation
12. Create educational resources





Content Source:
World Economic Forum (WEF)


Murat Durmus
(CEO AISOMA)
9 ethical issues in Artificial Intelligence
Unemployment
What happens after the end of jobs?
Inequality
How do we distribute the wealth created by machines?
Humanity
How do machines affect our behavior and interaction?
Evil genies
How do we protect against unintended consequences?
Artificial stupidity.
How can we guard against mistakes?
Racist robots.
How do we eliminate AI bias?
Security.
How do we keep AI safe from adversaries?
Singularity
How do we stay in control of a complex intelligent system?
Robot rights
How do we define the humane treatment of AI?
Content Source:
World Economic Forum (WEF)

Ethics : Opposition

AI should not replace

- Doctors
- Judges
- Police officers,etc

Some Personal Experiences

Many are concerned about
the lack of *AI-Experts*.
The lack of "real" *Thinkers* & *Philosophers*
is even more alarming.

Murat Durmus
(CEO AISOMA)

Related Publications

A. Tiwari, **S. Saha** , P. Bhattacharyya (2022), ``A Knowledge Infused Context Driven Dialogue Agent for Disease Diagnosis using Hierarchical Reinforcement Learning", **Knowledge Based Systems (impact factor:8.038)** .

A. Tiwari, T. Saha, **S. Saha** , S. Sengupta, A. Maitra, R. Ramnani, P. Bhattacharyya (2021), ``A Persona Aware Persuasive Dialogue Policy for Dynamic and Co-operative Goal Setting", **Expert Systems with Applications (impact factor: 6.954, h5-index: 118)**.

A. Tiwari, T. Saha, **S. Saha**, S. Sengupta, A. Maitra, R. Ramnani, and P. Bhattacharyya(2021), ``Multi-Modal Dialogue Policy Learning for Dynamic and Co-operative Goal Setting" in **International Joint Conference on Neural Networks (IJCNN) 2021**, 18-22 July 2021

T. Saha, D. Gupta, **S. Saha** , P. Bhattacharyya (2021), ``A Unified Dialogue Management Strategy for Multi-Intent Dialogue Conversations in Multiple Languages", **ACM Transactions on Asian and Low-Resource Language Information Processing**

A. Tiwari, T. Saha, **S. Saha**, S. Sengupta, A. Maitra, R. Ramnani, P. Bhattacharyya (2021), ``A Dynamic Goal Adapted Task Oriented Dialogue Agent", **Plos One (h5 index: 175, impact factor: 2.74)**

T. Saha, N. Priya, **S. Saha** and P. Bhattacharyya (2021), ``A Transformer based Multi-task Model for Domain Classification, Intent Detection, and Slot-Filling" in **International Joint Conference on Neural Networks (IJCNN) 2021**, 18-22 July 2021.

T. Saha, S. Chopra, **S. Saha**, P. Bhattacharyya and Dr. P. Kumar (2021), ``A Large-Scale Dataset for Motivational Dialogue System: An Application of Natural Language Generation to Mental Health", in **International Joint Conference on Neural Networks (IJCNN) 2021**, 18-22 July 2021

T. Saha, A. Upadhyaya, **S. Saha**, P. Bhattacharyya (2021), ``Towards Sentiment and Emotion aided Multi-modal Speech Act Classification in Twitter", in **NAACL-HLT 2021**, June 6-11, 2021

T. Saha, S. Chopra, **S. Saha** and P. Bhattacharyya (2020), ``Reinforcement learning based personalized neural response generation", in **International Conference on Neural Information Processing (ICONIP) 2020**, 18-22 November, 2020

N. Priya, A. Tiwari and **S. Saha**(2021), ``Context Aware Joint Modeling of Domain Classification, Intent Detection and Slot Filling with Zero-shot Intent Detection Approach", in **28th International Conference on Neural Information Processing (ICONIP-2021),**. December 8 - 12, 2021, BALI, Indonesia

T. Saha, A. Patra, **S. Saha** and P. Bhattacharyya (2020), `` Towards Emotion-aided Multi-modal Dialogue Act Classification", In **ACL 2020**, July 5-10, 2020, Seattle, Washington

References

- Li, X., Chen, Y. N., Li, L., Gao, J., & Celikyilmaz, A. (2017, November). End-to-End Task-Completion Neural Dialogue Systems. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 733-743).
- Shi, W., & Yu, Z. (2018, July). Sentiment Adaptive End-to-End Dialog Systems. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1509-1519).
- Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., & Yu, Z. (2019). Persuasion for good: Towards a personalized persuasive dialogue system for social good. arXiv preprint arXiv:1906.06725.
- Shi, W., Wang, X., Oh, Y. J., Zhang, J., Sahay, S., & Yu, Z. (2020, April). Effects of persuasive dialogues: testing bot identities and inquiry strategies. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-13).
- Tang, K. F., Kao, H. C., Chou, C. N., & Chang, E. Y. (2016, December). Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In NIPS Workshop on Deep Reinforcement Learning.
- Kao, H. C., Tang, K. F., & Chang, E. (2018, April). Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).
- Wei, Z., Liu, Q., Peng, B., Tou, H., Chen, T., Huang, X. J., ... & Dai, X. (2018, July). Task-oriented dialogue system for automatic diagnosis. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 201-207).
- Liao, K., Liu, Q., Wei, Z., Peng, B., Chen, Q., Sun, W., & Huang, X. (2020). Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning. arXiv preprint arXiv:2004.14254.
- Rasmussen, K., Belisario, J. M., Wark, P. A., Molina, J. A., Loong, S. L., Cotic, Z., ... & Car, J. (2014). Offline eLearning for undergraduates in health professions: a systematic review of the impact on knowledge, skills, attitudes and satisfaction. Journal of global health, 4(1).
- Ramakrishnan, N., Vijayaraghavan, B. K. T., & Venkataraman, R. (2020). Breaking barriers to reach farther: A call for urgent action on tele-ICU services. Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine, 24(6), 393.
- Fox, S., & Duggan, M. (2013). Health online 2013. Health, 2013, 1-55.
- Conversational AI Market by Component (Platform and Services), Type (IVA and Chatbots), Technology (ML and Deep Learning, NLP, and ASR), Application, Deployment Mode (Cloud and On-premises), Vertical, and Region - Global Forecast to 2025." Markets and Markets, June 2020. Retrieved from: https://www.researchanmarkets.com/reports/5136158/conversational-ai-market-by-component-platform?utm_sourc=GNOM&utm_medium=PressRelease&utm_code=m3d7t3&utm_campaign=1426757+-+Global+Conversational+AI+Market+Analysis+2020-2025&utm_exec=joca220prd

Thank You

For any further question/remark/suggestion, please contact
Dr. Sriparna Saha
Email Id : sriparna@iitp.ac.in

How to learn an Optimal Policy ?

❑ How good is a state (S) ?

Expected cumulative reward from state, *S*

Intitution :

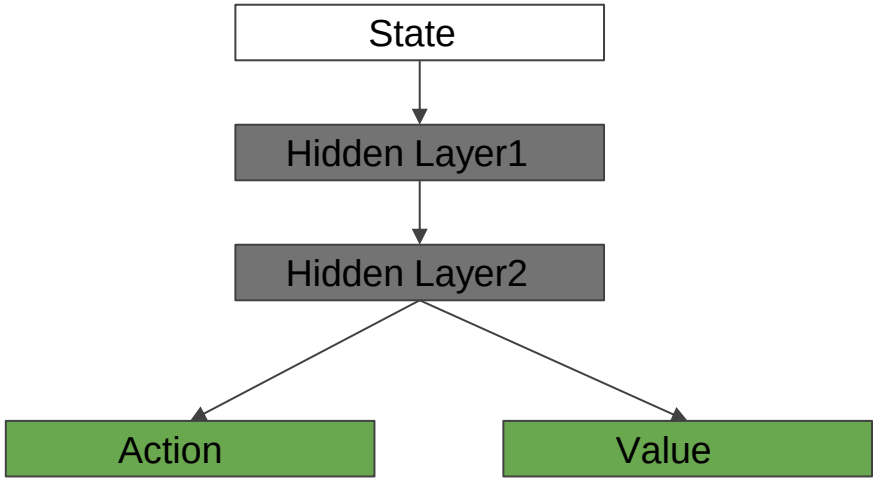
- *good state : If the state, S, is more likely to lead to a path that ends up to a goal state.*
- *bad state : If the state S, is more likely to lead to a path that would not be end up at one of goal states.*

How good or bad (intensity / subjectivity) ?

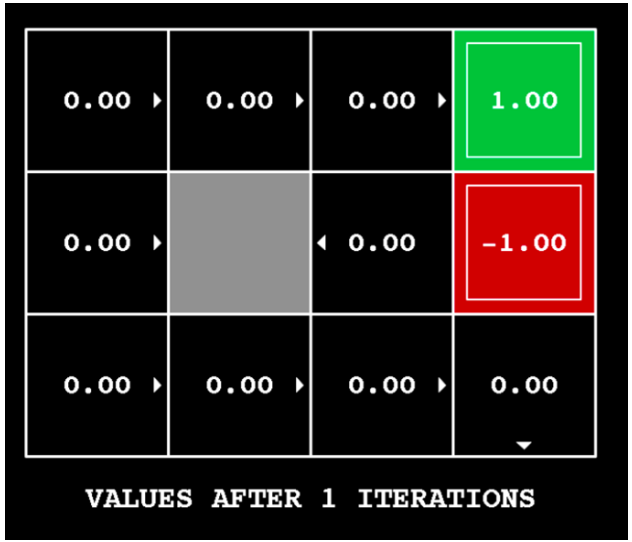
Value (S_{11}) = p (S_{11} → S_{21}) * 8 + p (S_{11} →S_{22}) * 2

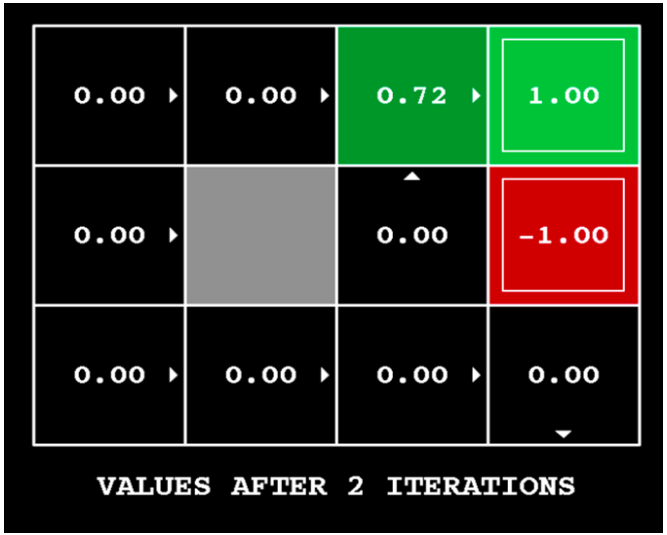
+ 50 (Assuming $\gamma = 1$)

= 0.9 * 8 + 0.1 * 2 + 50
= 57.4


State
Hidden Layer1
Hidden Layer2
Action
Value

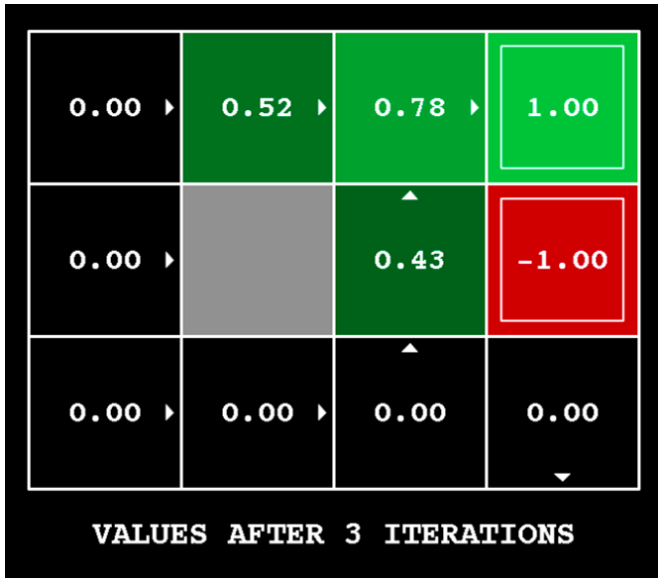
An Example

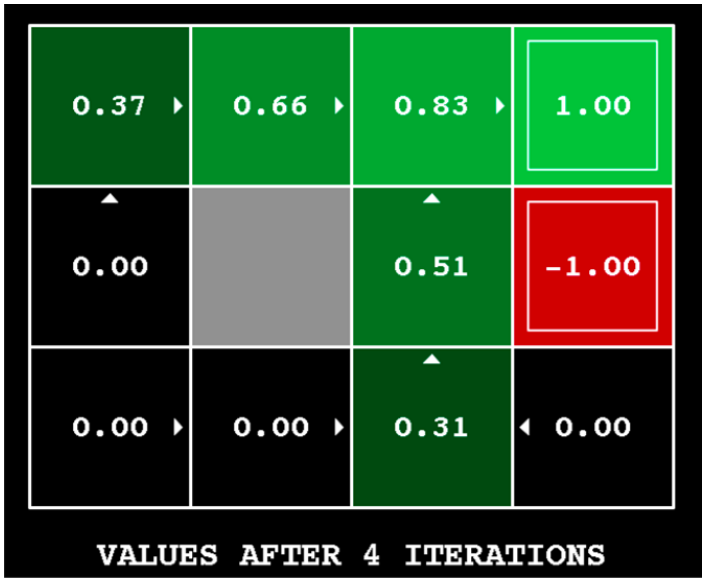




y =0.9, two terminal states with R = +1 and -1

An Example


0.00
0.52
0.78
1.00
0.00
0.43
-1.00
0.00
0.00
0.00
0.00
VALUES AFTER 3 ITERATIONS

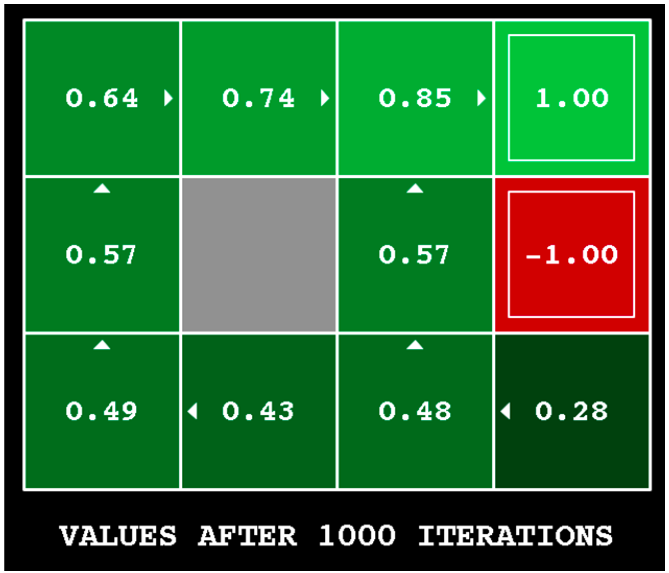

0.37
0.66
0.83
1.00
0.00
0.51
-1.00
0.00
0.00
0.31
0.00
VALUES AFTER 4 ITERATIONS

An Example


0.51
0.72
0.84
1.00
0.27
0.55
-1.00
0.00
0.22
0.37
0.13
VALUES AFTER 5 ITERATIONS


0.64
0.74
0.85
1.00
0.57
0.57
-1.00
0.49
0.43
0.48
0.28
VALUES AFTER 100 ITERATIONS

An example


0.64
0.74
0.85
1.00
0.57
0.57
-1.00
0.49
0.43
0.48
0.28
VALUES AFTER 1000 ITERATIONS