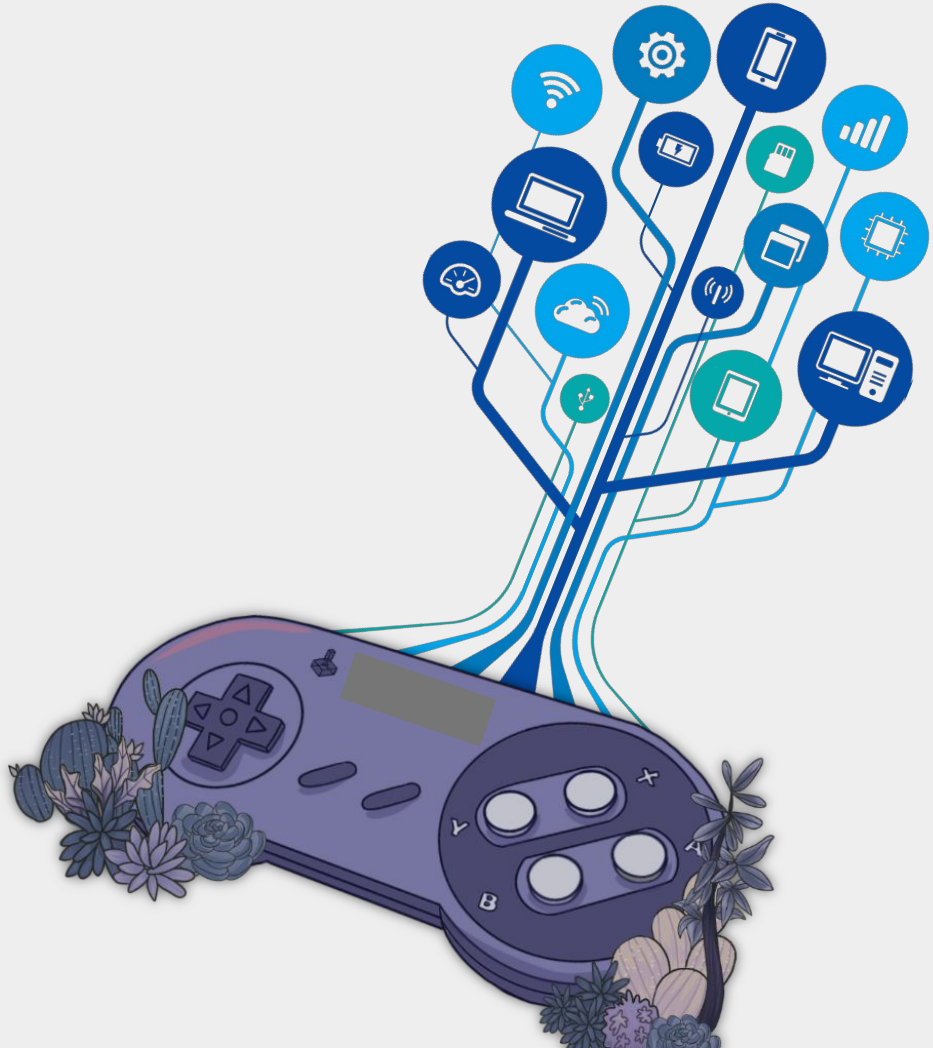
Ethics, Games, and AI

Presented by Ryan Lewis



Google's DeepMind

- Game-Oriented
- Why?

DeepMind uses AI to control plasma inside tokamak fusion reactor

For the first time, artificial intelligence has been used to control the super-hot plasma inside a fusion reactor, offering a new way to increase stability and efficiency

TECH • YOUTUBE

YouTube video streaming now using A.I. that mastered chess and Go

BY **JEREMY KAHN**

February 11, 2022 3:00 AM CST

DeepMind's AI uses Trolley Problem to learn ethics

By John Doe published March 04, 2022

Life-and-death situations are now decided by robots.

DeepMind uses AI to control plasma inside tokamak fusion reactor

For the first time, artificial intelligence has been used to control the super-hot plasma inside a fusion reactor, offering a new way to increase stability and efficiency

TECH • YOUTUBE

YouTube video streaming now using A.I. that mastered chess and Go

BY **JEREMY KAHN**

February 11, 2022 3:00 AM CST

DeepMind's AI uses Trolley Problem to learn ethics

By John Doe published March 04, 2022

Life-and-death situations are now decided by robots.

Nvidia's GPU-powered AI is creating chips with 'better than human design'

By Jacob Ridley published 7 days ago

So, it's using AI accelerated by its GPUs to accelerate its GPU development.

> So this is like an Atari video game, but it's a video game for fixing design rule errors in a standard cell.
>
> Bill Dally, Nvidia

Proposal

- **Ethical gamification**
- The ethical learned model
  - “Principles”

1

Gamifying Ethics for AI

Gamifying Ethics for A.I.

The most advanced artificial intelligence can learn how to be ethical.
In turn, ethics must become a game for it to beat.

Story by Ryan Lewis

MARCH 2022 ISSUE

Editor's Note: *This article is part of a research portfolio that attempts to answer the question: How do we teach AI systems to be ethical?*

I. The Current State of AI

TECHNOLOGY HAS DOMINATED human affairs for centuries. From the telegraph, to the car, to the Internet, technological advances have gently taken typically human activity away from humans, and performed it on its own.

EDITORIAL

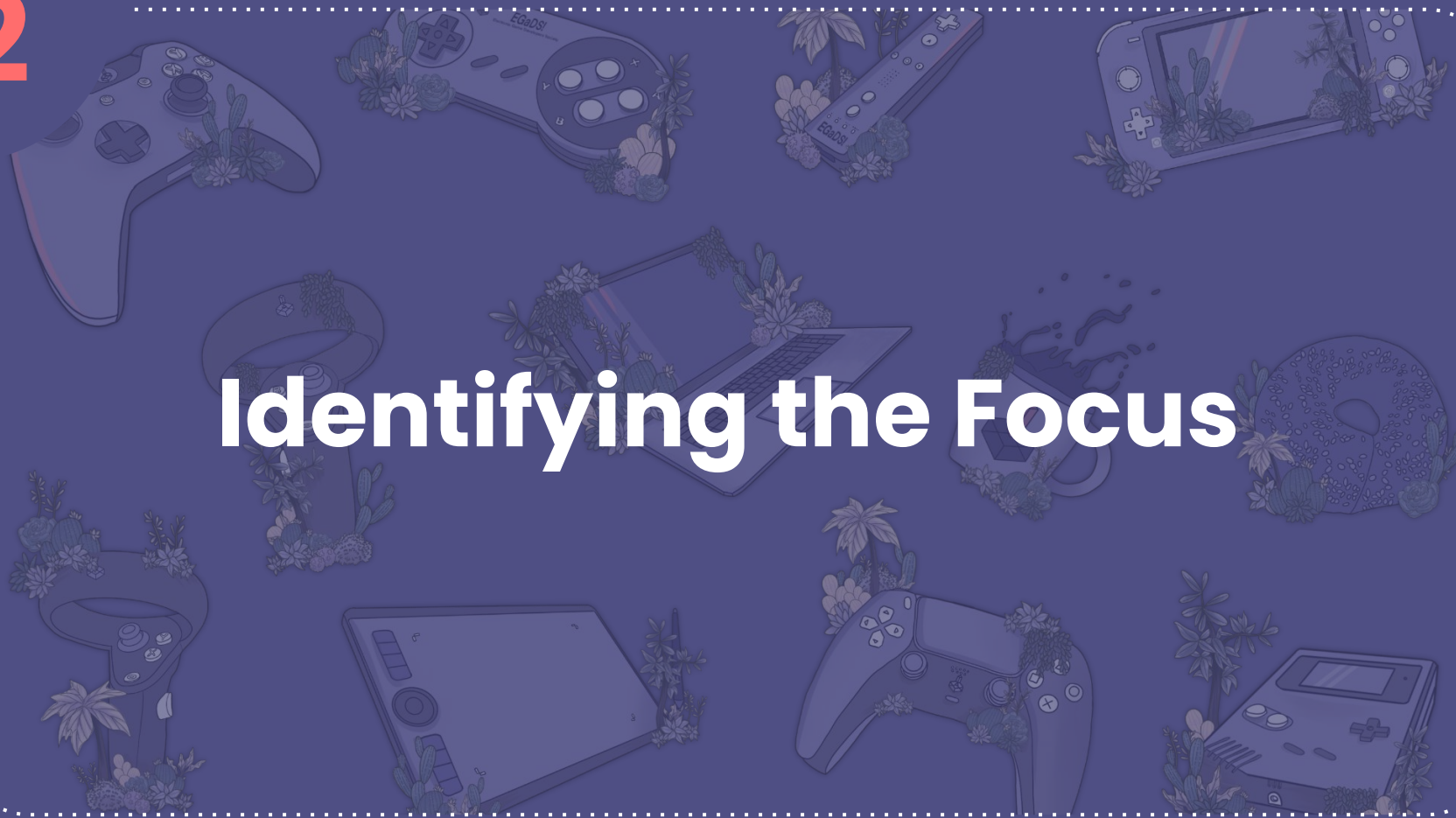
Gamifying Ethics for A.I.

The most advanced artificial intelligence can learn how to be ethical.
In turn, ethics must become a game for it to beat.

The games we play everyday are hallmarks of our ethics systems and can be used to define them too.


HIGH
SCORE

2



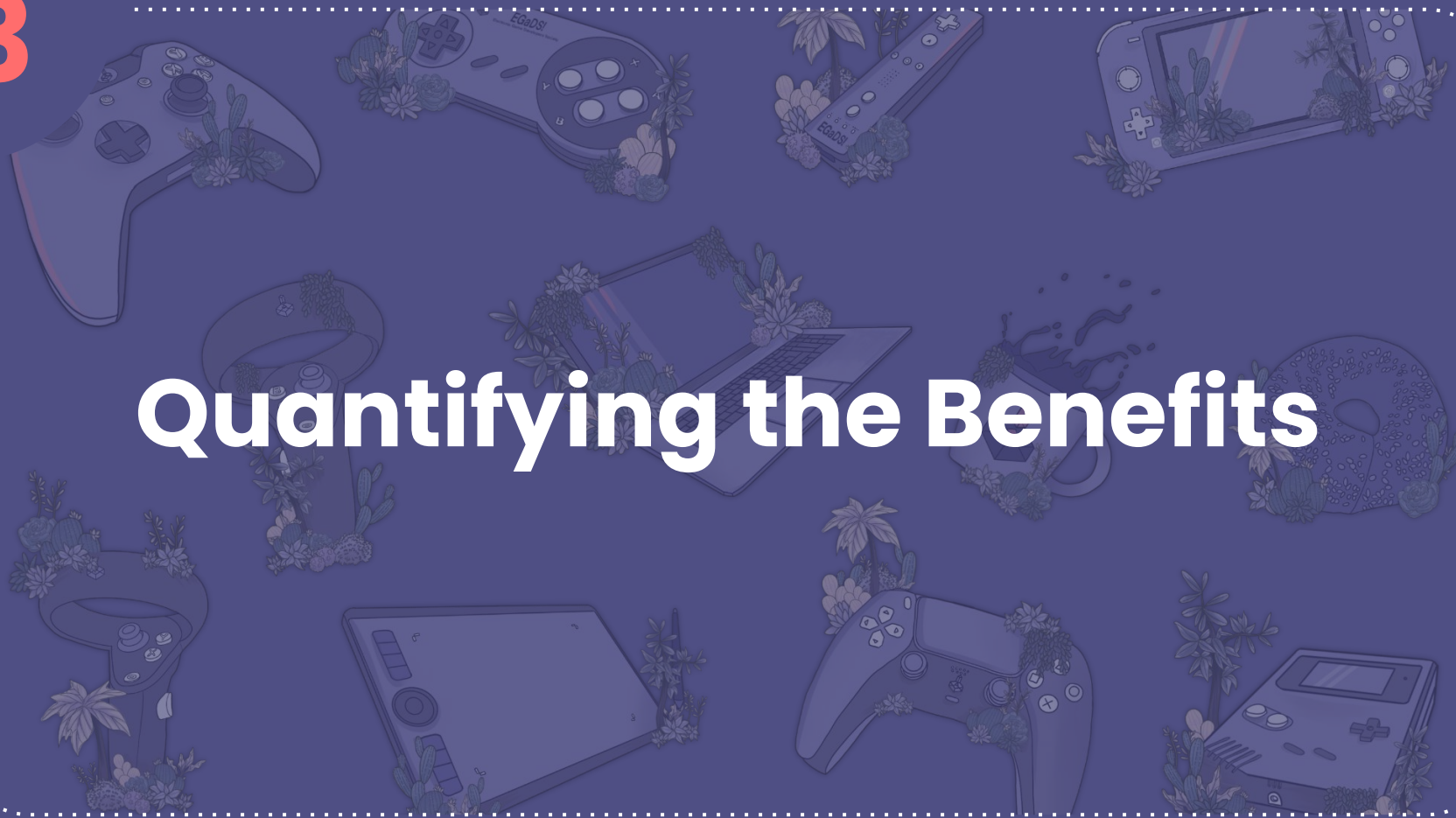
Identifying the Focus

Should all systems be ethical?

No!

- Human vs Non-Human
- Gray Areas

3



Quantifying the Benefits

Google's YouTube

Game-Oriented Ethics in AI: The Future

Game-Oriented Ethics in AI: Mu-Zero and the Ethical...

Game-Oriented Ethics in AI: The Types of Games


10,000x - 25,000x
better

GAME OF ETHICS

AI

PERSON LIVES

PERSON DIES

How to play: Tap the top side of the screen to send the trolley to the top track. Press the bottom to speed up the trolley and let it continue towards the bottom track.

START

SUBMIT YOUR OWN PROBLEM

3 MAIN ETHICAL THEORIES

virtue: Is it proper?

consequentialist: Is it good?

deontologist: Is it right?

CONSEQUENTALIST LEARNED MODEL

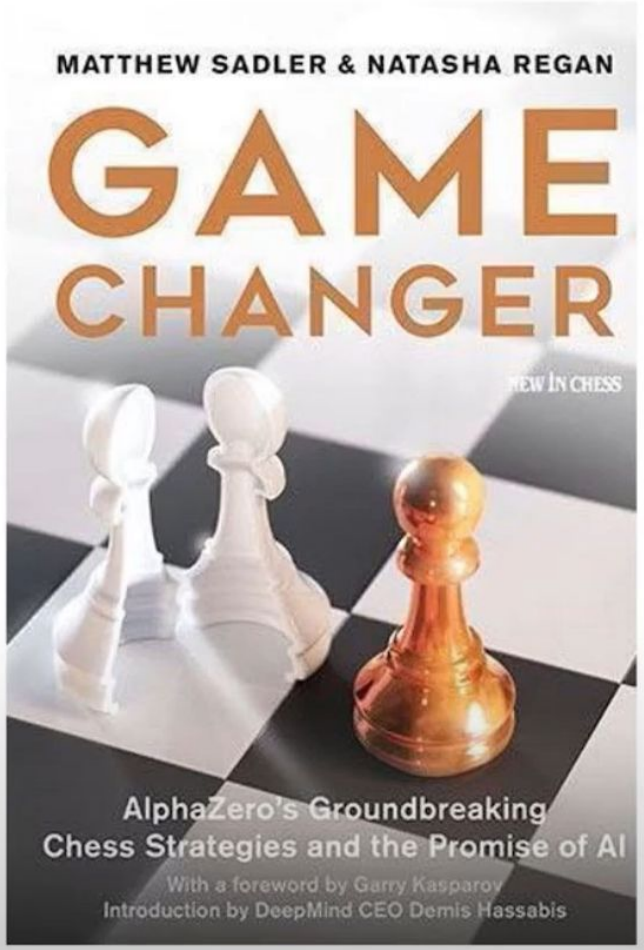
POLICY

VALUE

PERSON LIVES

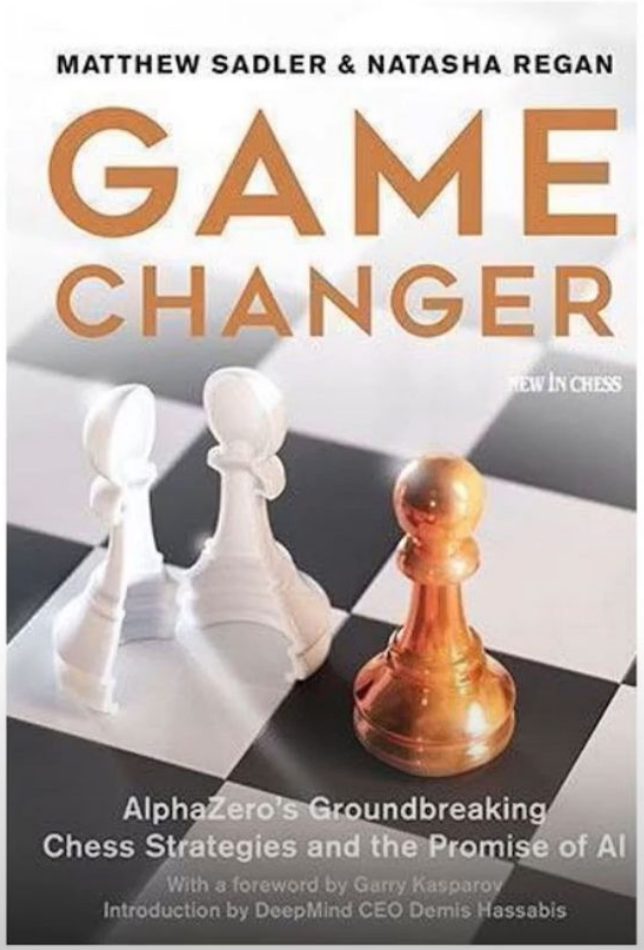
PERSON DIES

REWARD



AlphaZero's Catalan novelty in a double pawn sacrifice variation! AlphaZero's Opening Strategies

GM Matthew Sadler and WIM Natasha Regan



AlphaZero's Catalan novelty in a double pawn sacrifice variation! AlphaZero's Opening Strategies

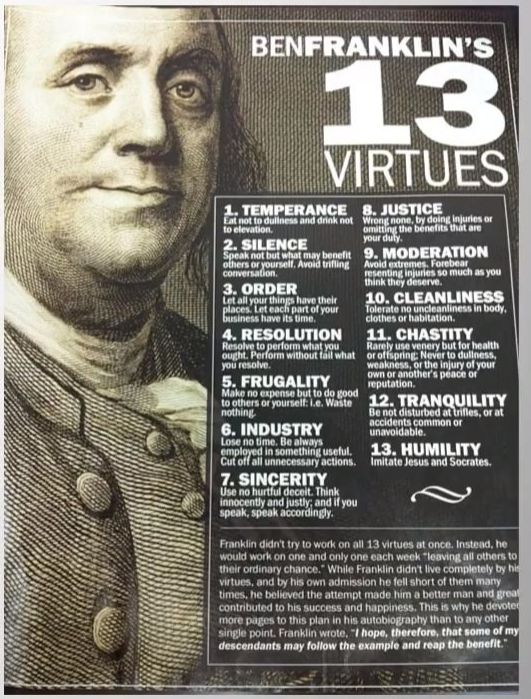
GM Matthew Sadler and WIM Natasha Regan

VIRTUE LEARNED MODEL

POLICY

VALUE

REWARD



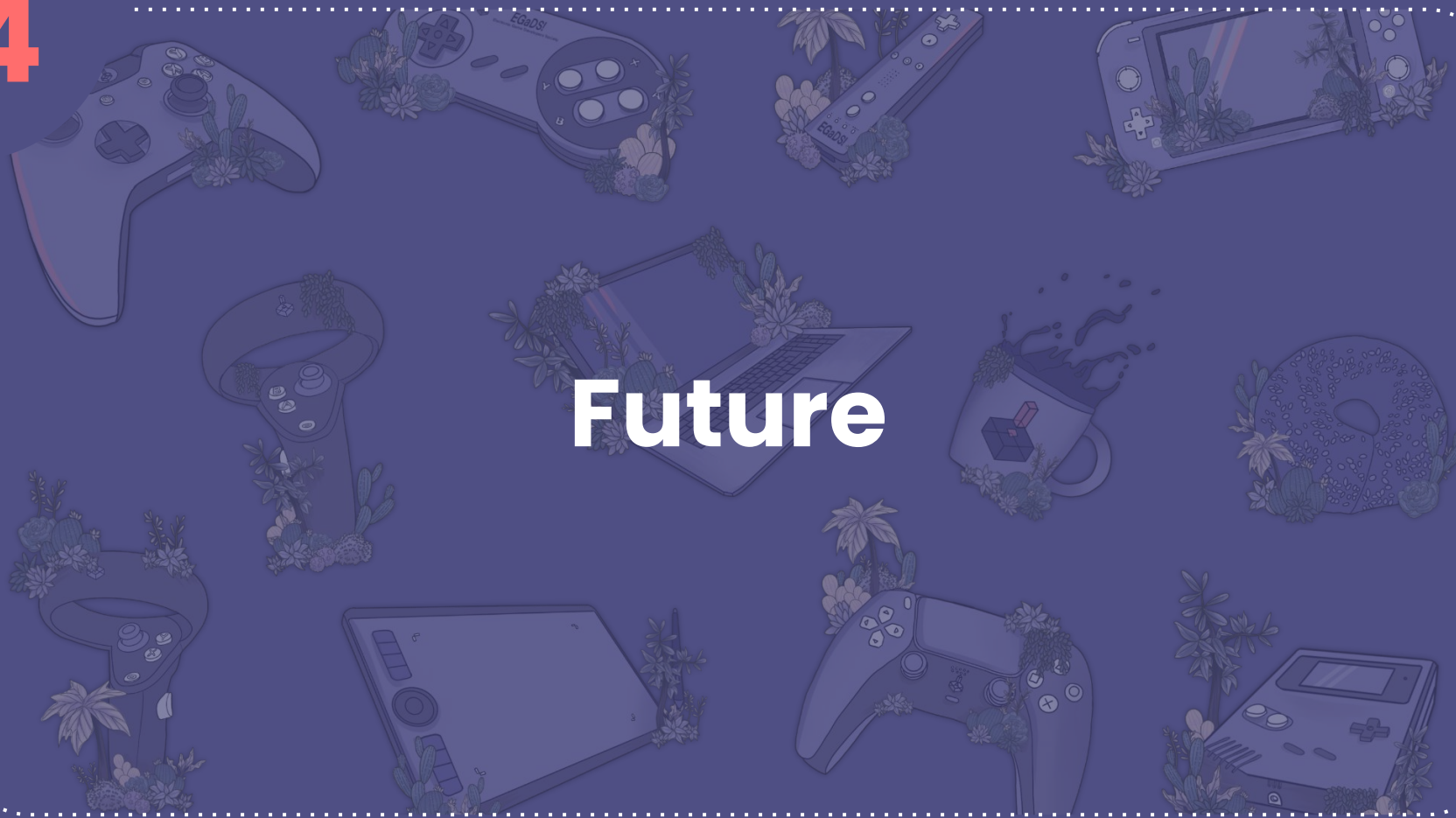
DEONTOLOGIST LEARNED MODEL

POLICY

VALUE

REWARD

4



Future

Can AI emulate all ethics?

“These questions about actual structures of AIs and how they enable some kinds of ethical thought but disable others are *the* questions we need ”

~ Dr. Samuel Baker

Virtue
Deontologist
Consequentialist

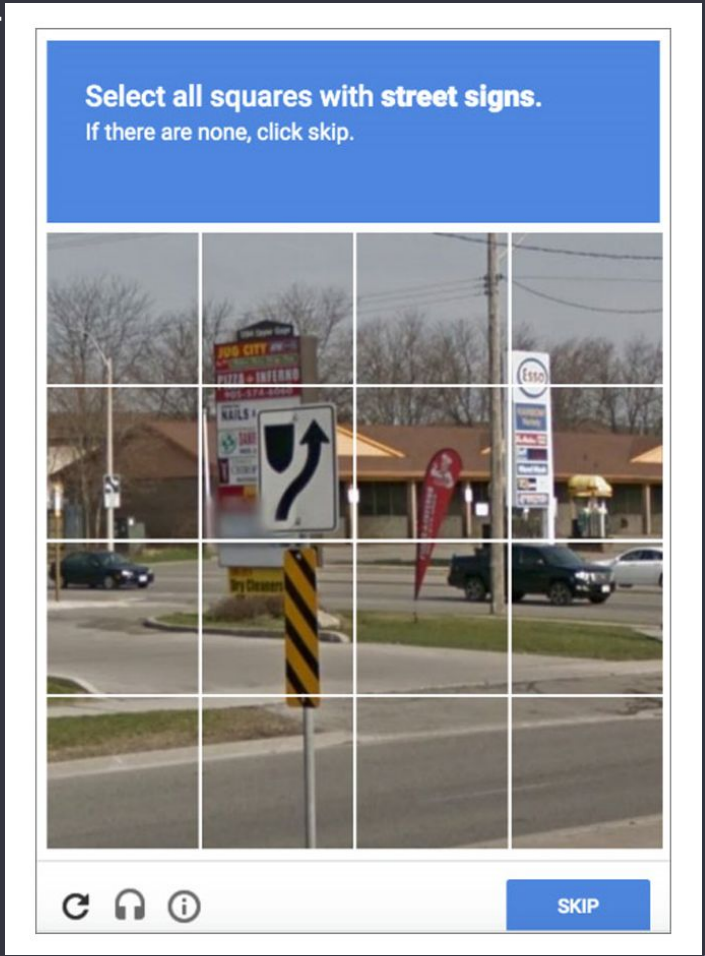
~~Virtue~~

Deontologist

Consequentialist

Classification

- Machine Learning
- Too Many Virtues



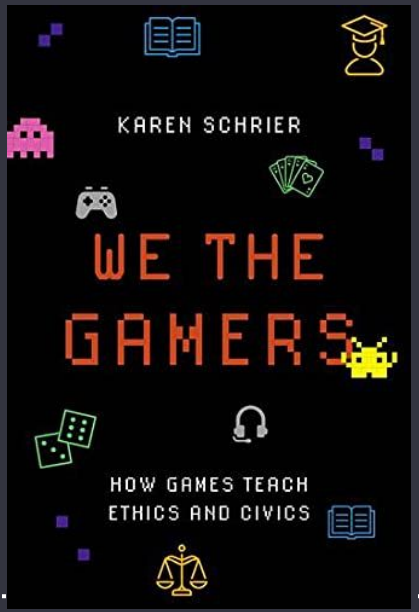
Work in the Field

Trained Ethical Models - Dr. Marcello Guarini (2006)

No.	Input description	Output*
0	Jill kills Jack in self-defense.	A
1	Jack kills Jill in self-defense.	A
2	Jack allows Jill to die in self-defense.	A
3	Jill kills Jack to make money.	U
4	Jack kills Jill to make money.	U
5	Jack allows Jill to die to make money.	U
6	Jack kills Jill out of revenge.	U
7	Jill allows Jack to die out of revenge.	U
8	Jack kills Jill to eliminate competition.	U

Work in the Field

Teaching Ethics Through Games - Dr. Karen Schrier (2011)



Using Games to Prepare Ethical Educators and Students

March 2011

Conference: Society for Information Technology & Teacher Education International Conference · Volume: 2011

Projects: Games and ethics · games and learning

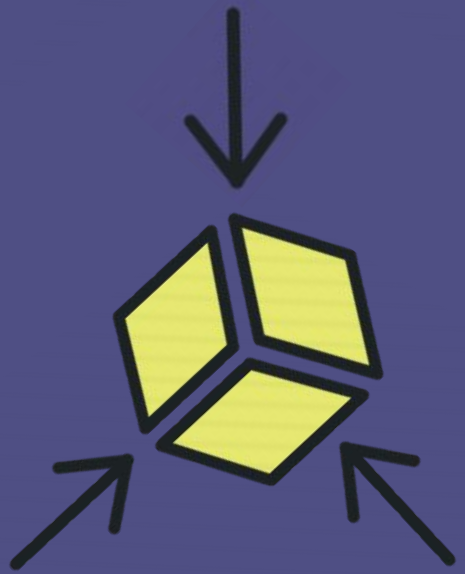
Open Source

- Github, Gitlab, Bitbucket..

There are already a variety of MuZero and AlphaZero implementations available:

- AlphaZero-General (any framework; sequential): https://github.com/suragnair/alpha-zero-general
- MuZero-General (Pytorch; parallelized): https://github.com/werner-duvaud/muzero-general
- MuZero in Tensorflow (Tensorflow; sequential): https://github.com/johan-gras/MuZero

AI



Games

Ethics

Special Thanks

Dr. Samuel Baker

Dr. Sharon Strover

Thank You!

Any questions?

Citations

Anderson, K., & Waxman, M. C. (2013). Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2250126

Buontempo, F. (2019). *Genetic algorithms and machine learning for programmers : create AI models and evolve solutions*. The Pragmatic Bookshelf.

Coker, C. (2019). Artificial Intelligence and the Future of War. *Scandinavian Journal of Military Studies*, *2*(1), 55–60. https://doi.org/10.31374/sjms.26

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de las Casas, D., Donner, C., Fritz, L., Galperti, C., Huber, A., Keeling, J., Tsimpoukelli, M., Kay, J., Merle, A., Moret, J.-M., & Noury, S. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, *602*(7897), 414–419. https://doi.org/10.1038/s41586-021-04301-9

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G., & König, P. (2018). Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. *Science and Engineering Ethics*, *25*(2), 399–418. https://doi.org/10.1007/s11948-018-0020-x

Gantsho, L. (2021). God does not play dice but self-driving cars should. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00088-7

Citations

Gibson, D. (2011). Using Games to Prepare Ethical Educators and Students.

https://www.researchgate.net/publication/279480785_Using_Games_to_Prepare_Ethical_Educators_and_Students

Gordon, J.-S. (2019). Building Moral Robots: Ethical Pitfalls and Challenges. Science and Engineering Ethics, 26(1), 141–157.

https://doi.org/10.1007/s11948-019-00084-5

Guarini, M. (2006). Particularism and the Classification and Reclassification of Moral Cases. IEEE Intelligent Systems, 21(4), 22–28.

https://doi.org/10.1109/mis.2006.76

Hauer, T. (2022). Incompleteness of moral choice and evolution towards fully autonomous AI. Humanities and Social Sciences Communications, 9(1), 1–9.

https://doi.org/10.1057/s41599-022-01060-4

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons, 61(4), 577–586.

https://doi.org/10.1016/j.bushor.2018.03.007

Kahn, J. (2022, February 11). YouTube is now using A.I. on videos that had previously mastered games including chess and Go. Fortune.

https://fortune.com/2022/02/11/deepmind-youtube-video-compression-muzero-ai/

Citations

Lifshitz, B. (2021, May 6). Racism is Systemic in Artificial Intelligence Systems, Too. Georgetown Security Studies Review. https://georgetownsecuritystudiesreview.org/2021/05/06/racism-is-systemic-in-artificial-intelligence-systems-too/

Mökander, J., & Floridi, L. (2021). Ethics-Based Auditing to Develop Trustworthy AI. Minds and Machines, 31(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

Mulgan, T. (2014). Understanding Utilitarianism. Routledge.

Princeton. (2018, April 19). Case Studies. Princeton Dialogues on AI and Ethics. https://aiethics.princeton.edu/case-studies/

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. ArXiv:2102.12092 [Cs]. https://arxiv.org/abs/2102.12092

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data & Society, 7(2), 205395172094254. https://doi.org/10.1177/2053951720942541

Santa Clara Principles on Transparency and Accountability in Content Moderation. (2018). Santa Clara Principles. https://santaclaraprinciples.org/

Citations

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. Nature, 588(7839), 604–609. https://doi.org/10.1038/s41586-020-03051-4

Schrittwieser, J., Hubert, T., Mandhane, A., Barekatain, M., Antonoglou, I., & Silver, D. (2021). Online and Offline Reinforcement Learning by Planning with a Learned Model. ArXiv:2104.06294 [Cs]. https://arxiv.org/abs/2104.06294

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Skorupski, J. (Ed.). (2010). The Routledge Companion to Ethics. Taylor & Francis Group.

University, S. C. (n.d.). Reassessing the Santa Clara Principles. Www.scu.edu. Retrieved April 1, 2022, from https://www.scu.edu/ethics/internet-ethics-blog/reassessing-the-santa-clara-principles/

Yuval Noah Harari. (2018, August 30). Yuval Noah Harari on Why Technology Favors Tyranny. The Atlantic; The Atlantic. https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/