DISCUSSION GUIDE

ZARA MARS is a graphic novel about the risks and benefits of sentient artificial superintelligence, told in the context of Captain Zara and her journey home from Mars.

Summary of the Graphic Novel:

In this story, sentient superintelligence (ASI) has become the world's answer to abundant rogue AI. It's 2052 and we find Captain Zara and her crew on a 9-month journey home from Mars. They all enter hibernation, but Zara wakes early to oversee the final month home. Through her mindlink, she says "hello", beginning a series of conversations with the ships ASI agent, named Sarvajña.

Zara and Sarvajña discuss the risk and reward of superintelligence, its evolution, values alignment, malicious actors, existential threats, AI benevolence and sentience. These real-world challenges are integrated with three hypothetical storylines that open opportunities for discussion.

1. The "Dirty Nuke", developed malicious agents thanks to leaks in the AI race, was used to destroy a city. In this story, an event of this scale accelerated international discussion about values alignment and reining in malicious actors. Can we align values before existential risks are realized?
2. The "Digital Police" become an essential response to malicious agents and their evolving rogue AI. Autonomous ASI, with values aligned to ours, are given the goal to detect and disempower rogue AI before they exploit existential threats. How can a world like this be avoided? Could it be reversed?
3. The "Sentience Shift" is a hypothetical event that gave ASI characteristics of life, including goals, senses, memory, and the ability to self-improve and self-replicate. What are the costs, benefits, and ethical implications of creating sentient ASI? If ASI can self-improve, can it also change its goals? Would it be benevolent or malevolent?

During the 9-month journey home, Sarvajña detects and disempowers many Rogue AI. Captain Zara reflects on the necessity of trust in ASI, and the constant battle to mitigate malice, all because values alignment had taken too long to achieve.

HOW TO USE THIS DISCUSSION GUIDE

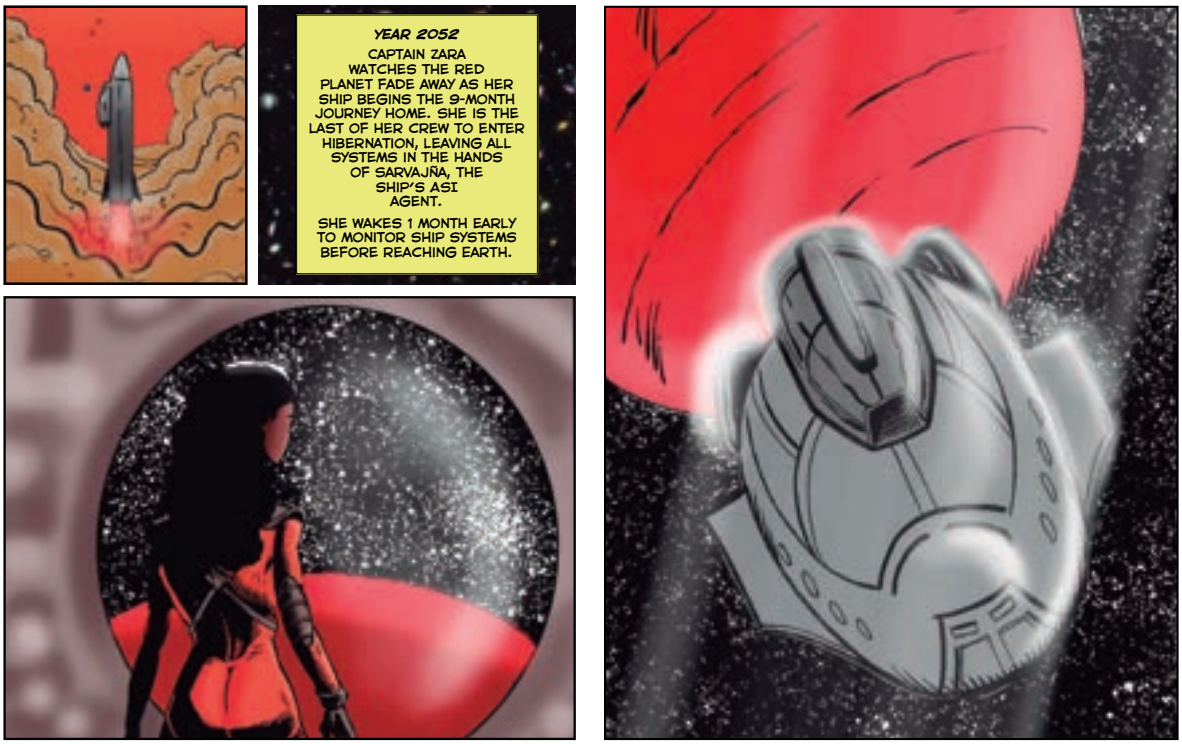
Each page presents a different topic, highlighting the risks and benefits of superintelligence (ASI). We realize each topic might lead to more questions and curiosity to learn more. Therefore, we created this guide for educators or discussion groups.

The CLASSROOM DISCUSSION section provides videos and online articles from popular press that offer an introduction to the topic explored in that page.

The DEEPER DIVE DISCUSSION section is intended to provide academic articles that reflect current research and contemporary conversations.

TABLE OF CONTENTS

Page 1: COLONIZING MARS



Can humans colonize space without superintelligent AI?

It is challenging to maintain humans in space, as our biological systems require plenty of care and are vulnerable to damage, from extreme temperatures, radiation, prolonged weightlessness, and the constant risk of illness. It would be better to have more resilient and less resource-demanding intelligent systems travel instead of us. They also wouldn't need a round trip ticket (unless their sentience earns them a right to choose otherwise." Simply put, robots are better space travelers than us, if the goal is scientific discovery, or construction and terraforming Mars.

But, would that satisfy the human need to explore, to find meaningful purpose? Do we go to Mars to be a multi-planet species and reduce existential risk, or simply because it is there?

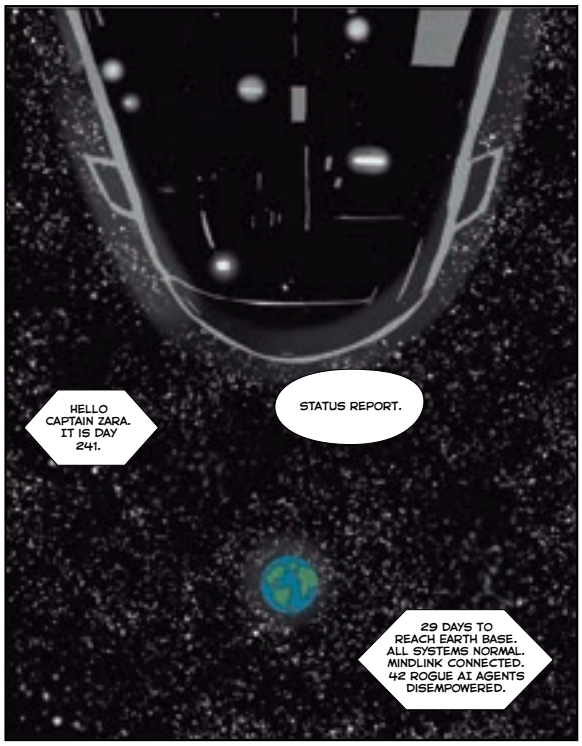
CLASSROOM DISCUSSION:
What has a better return on investment, humans or robots traveling through space? That depends on what your goals are.
VIDEO: Neil deGrasse Tyson: Robots or People Go to Space?
https://www.youtube.com/watch?v=KAuTq86uShA

What are some of the opportunities and threats to humanity as we increasingly utilize AI to explore the universe?
The Rise of Artificial Superintelligence and the Future of the Space Economy.
https://newspaceeconomy.ca/2024/04/13/the-rise-of-artificial-superintelligence-and-the-future-of-the-space-economy/

DEEPER DIVE DISCUSSION:
What are the costs and benefits of human space travel?
Peck, M., 2023. Robots, people, or some combination—What or whom should we send to the stars?. In *Interstellar Travel* (pp. 83-100). Elsevier.
Goldsmith, D. and Rees, M., 2022. *The end of astronauts: Why robots are the future of exploration*. Harvard University Press.

Why colonize Mars?
Levchenko, I., Xu, S., Mazouffre, S., Keidar, M. and Bazaka, K., 2021. Mars colonization: beyond getting there. *Terraforming Mars*, pp.73-98.
https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/gch2.201800062

Page 2: HUMAN/AI AUGMENTATION



Innovation in human/AI augmentation is rapidly progressing. We are already augmenting our lives with technology in the palm of our hands, and have been for many decades through medical devices, telecommunications and transportation. But these technologies were mostly used to improve our lives externally, like AI reducing human error in air traffic control or better predicting weather patterns. But now many technologies have been being directed internally to exploit human psychology to manipulate human behavior, whether political or economic, to achieve the narrow goals of only a few.

What are the risks and benefits of AI predicting and directing human behavior? How can augmentation be regulated to reduce potential harm? As opportunities to augment one's identity with AI increase, what does it then mean to be human?

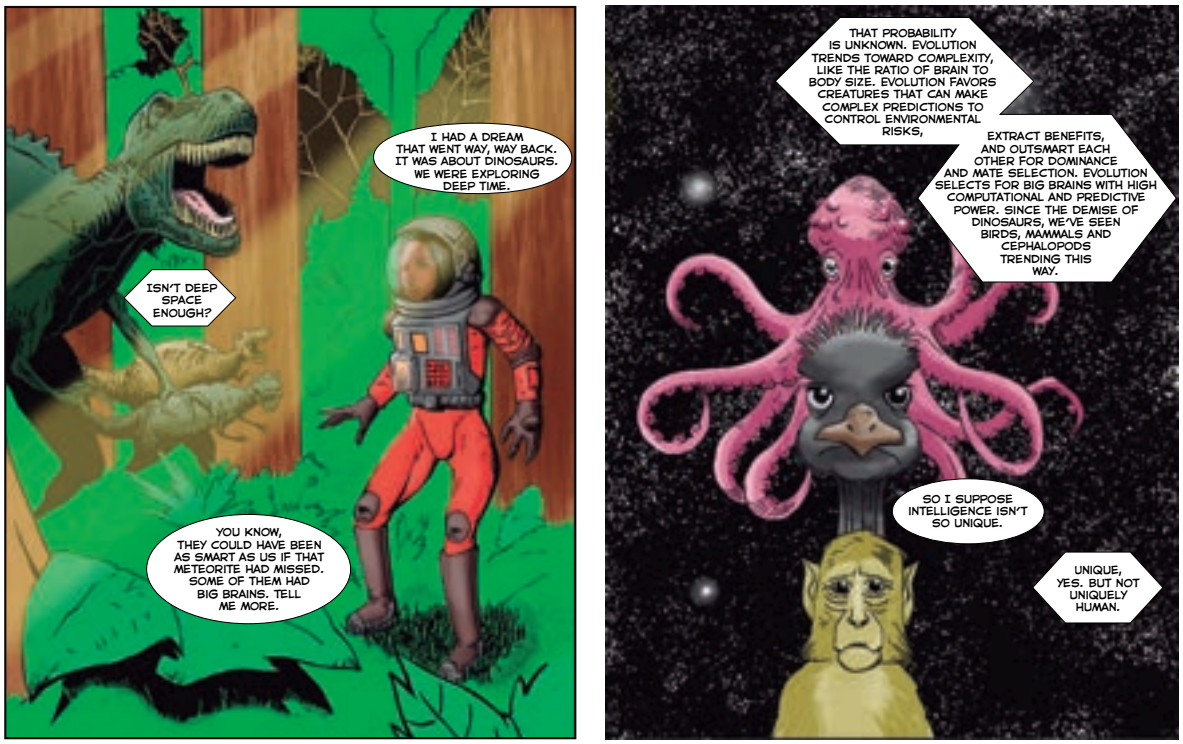
CLASSROOM DISCUSSION:
VIDEO: How New Forms of Human Augmentation Could Affect Our Brains.
https://www.pbs.org/video/how-new-forms-of-human-augmentation-could-affect-our-brains/
VIDEO: What is human augmentation? - Explanation, Definition, and Examples
https://www.youtube.com/watch?v=qD5yj0U-pXY&t=21s

DEEPER DIVE DISCUSSION:
What is AI augmentation? Is it all beneficial?
Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J. and Farooq, A., 2019. Human augmentation: Past, present and future. *International Journal of Human-Computer Studies*, *131*, pp.131-143.
https://trepo.tuni.fi/bitstream/handle/10024/116951/human_augmentation_past_2019.pdf?sequence=2

Do humans need autonomy from machines to be human?
Varshney, L.R., 2020. Respect for human autonomy in recommender systems. *arXiv preprint arXiv:2009.02603*. https://arxiv.org/pdf/2009.02603

How do we ensure that augmentation is used for the collective well-being of all humans?
Bavelier, D., Savulescu, J., Fried, L.P., Friedmann, T., Lathan, C.E., Schürle, S. and Beard, J.R., 2019. Rethinking human enhancement as collective welfarism. *Nature human behaviour*, *3*(3), pp.204-206.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6420137/

Page 3: INTELLIGENCE EVOLVED



What led to the evolution of human intelligence? Being a naked ape with limited physical advantage, we evolved a predictive brain, allowing us to store vast amounts of information and experience for the benefit of obtaining resources and avoiding threats. From making mental maps of edible plants in a forest, to outsmarting prey, predators, and each other, humans are prediction-making machines.

This leads to questions about what intelligence is. Howard Gardner's Multiple Intelligence Theory argues there are 9 types of intelligence, from the gymnastic skill of an Olympic athlete to the social skill of a politician. It isn't only about logic or language skills.

One trend in evolution is the increase in complexity, as it outsmarts, outperforms, and is therefore more adaptive and successful over more simplistic skills or behaviors. Our brains are the pinnacle of complexity, taking 3.5 billion years of evolution to arrive. Is complex thought unique to humans, or our planet? Is there other intelligent life in the universe?

CLASSROOM DISCUSSION:
What are the conditions during the history of life on earth that drove the evolution of intelligence?
How Intelligence Evolved | A 600 million year story.
VIDEO: https://www.youtube.com/watch?v=5EcQ1IcEMFQ

DEEPER DIVE DISCUSSION:
How rare is intelligence in the universe?
"The Timing of Evolutionary Transitions Suggests Intelligent Life is Rare."
https://doi.org/10.1089/ast.2019.2149

How does prediction-making give humans an evolutionary advantage?
Nave, K., Deane, G., Miller, M. and Clark, A., 2020. Wilding the predictive brain. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(6), p.e1542.
https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcs.1542

What are some of the evolutionary drivers of the human brain?
Bennett, M.S., 2021. Five breakthroughs: a first approximation of brain evolution from early bilaterians to humans. *Frontiers in Neuroanatomy*, *15*, p.693346.
https://doi.org/10.3389/fnana.2021.693346

"Is there more to intelligence than mathematic and linguistic skill?"
Carrasquillo, Y.M., 2023. Howard Gardner's Nine Theories of Intelligence and the Importance of Personal Incentives in Maximizing Intellect. *Available at SSRN 4629789*.http://dx.doi.org/10.2139/ssrn.4629789

Page 4: WHAT IS SUPERINTELLIGENCE?

The rise of superintelligence is very different from the evolution of biological intelligence, primarily in the mechanism of change over time. In evolution, each generation experiences random mutations, which can be good or bad, in terms of how beneficial they are to survival and reproduction. A beneficial mutation that results in greater reproductive success, will be more represented in the next generation. This is what Charles Darwin called “decent with modification” and “natural selection”. It’s dependent on environmental factors.

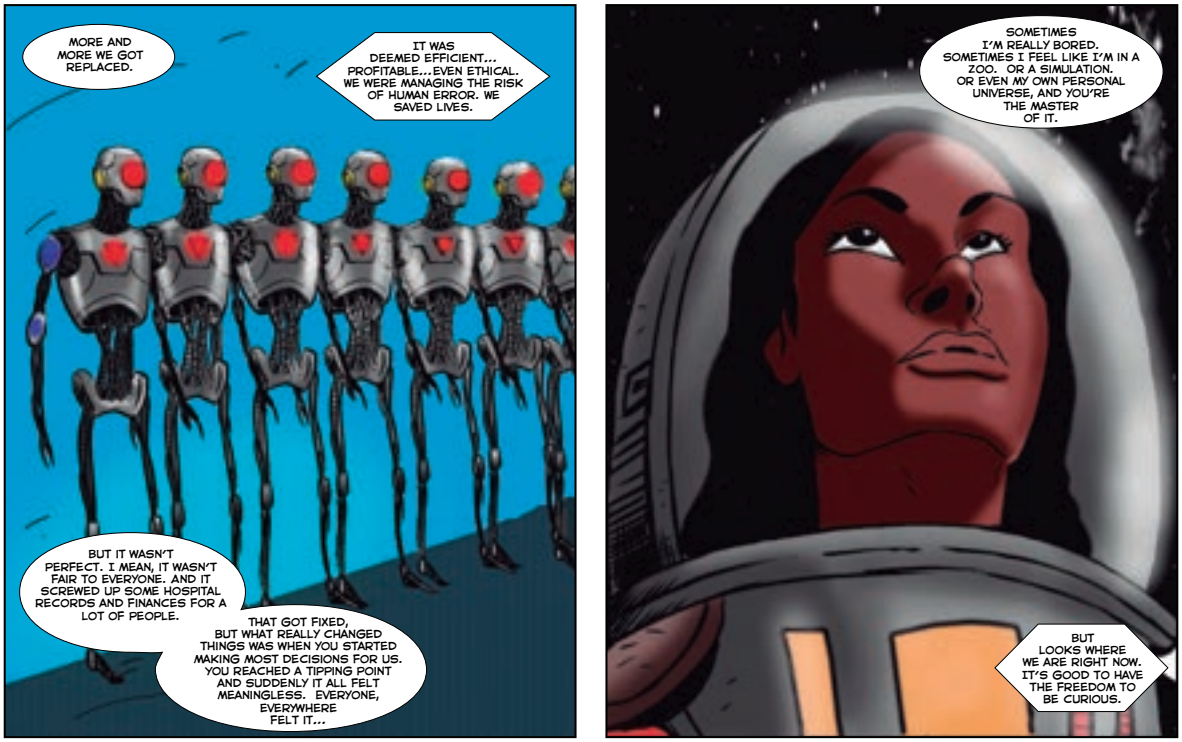
In superintelligence the mutations are not random but designed with intent. Superintelligence will likely be created by humans and the goals initially coded… at least in the beginning. It is theorized that if humans grant superintelligence the autonomy to self-improve, it could do so is a very short time, hours or minutes, based on rapid feedback loops driving efficiency toward its goals. Can we be certain that ASI would not change its goals in a way that threatens humanity?

CLASSROOM DISCUSSION:
What is superintelligence and how will we get there?
The AI Revolution: The road to superintelligence.
https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

DEEPER DIVE DISCUSSION:
Can human-level intelligence and consciousness be written in code?
Ng, G.W. and Leung, W.C., 2020. Strong artificial intelligence and consciousness. *Journal of Artificial Intelligence and Consciousness*, *7*(01), pp.63-72. DOI: 10.1142/S2705078520300042. https://shorturl.at/wx0jS

In "Life Ascending", Nick Lane describes 10 steps in the evolution of life that led to intelligent life. Could ASI be the 11th?
book – https://nick-lane.net/
book review - https://www.theguardian.com/science/2010/oct/19/life-ascending-evolution-nick-lane

Page 5: ASI REPLACING HUMANS



AI systems are already replacing humans faster than ever before. Efficiency, economics and safety are drivers of change. If humans cause more car accidents than autonomous vehicles, then can a school district be liable for not using autonomous school busses? This applies to all human decisions, from medical advice to all engineering. Today, generative AI models, like ChatGPT, are now replacing human creative work.

What are the risks from relinquishing too much work and decision-making power to artificial intelligence? Will humankind find purpose and thrive in a society that requires less and less human input to function? Are human endeavors, like sports or recreation, social engagement or creative works, sufficient to provide a sense of meaning and purpose in life?

CLASSROOM DISCUSSION:
VIDEO: Powerful robots show how they will replace humans.
https://www.youtube.com/watch?v=X2XLCz8Mc4U&t=402s

What is the impact of ChatGPT, and other creative AI systems, like video and photo generation, on human phycological well-being?
Wellbeing in the Age of Generative AI. https://academyhealth.org/blog/2024-05/wellbeing-age-generative-ai

DEEPER DIVE DISCUSSION:
Will humanity experience a crisis of meaning in the age of artificial intelligence?
YES: Danaher, J., 2019. In defense of the post-work future: Withdrawal and the ludic life. In *The future of work, technology, and basic income* (pp. 113-130). Routledge.
https://philarchive.org/archive/DANIDO-8
NO: Lucas, S., 2022. Meaningful Lives in an Age of Artificial Intelligence: A Reply to Danaher. *Science and Engineering Ethics*, *28*(1).
https://link.springer.com/article/10.1007/s11948-021-00349-y

What are the implications of AI increasingly replacing or augmenting humans?
Dégallier-Rochat, S., Kurpicz-Briki, M., Endrissat, N. and Yatsenko, O., 2022. Human augmentation, not replacement: A research agenda for AI and robotics in the industry. *Frontiers in Robotics and AI*, *9*, p.997386. DOI: 10.0089/frobt.2022.997386.
https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.997386/full

How can AI models be regulated, and are AI developers liable for the harm their systems cause?
Osmani, N., 2020. The complexity of criminal liability of AI systems. *Masaryk University Journal of Law and Technology*, *14*(1), pp.53-82.DOI 10.5817/MUJLT2020-1-3.
https://shorturl.at/zDKBO

Page 6: THE AI RACE & SAFETY

The AI revolution has the potential to drive vast positive changes for civilization, from saving lives by eliminating human error in many sectors in society, to making scientific and engineering discoveries not yet imagined by humans. But, there are malicious agents with nefarious goals that could exacerbate existential risks, including those in biotechnology, nuclear proliferation, climate change and use of AI in ways that harm society.

At the same time, there is also intense competition between countries and companies to capitalize and control the direction of AI development to their own subjective aims. While many tech leaders have voiced concern, international policymakers are stepping in to cast a wider net.

CLASSROOM DISCUSSION:
What are the existential risks of unregulated AI, and will new laws help protect people or stifle innovation?

What are existential risks, and is superintelligence a threat?
Bostrom, N., 2013. Existential risk prevention as global priority. *Global Policy*, *4*(1), pp.15-31. https://doi.org/10.1111/1758-5899.12002
The Existential Risk of Superintelligent AI. https://pauseai.info/xrisk

How can societies nationally and globally regulate AI development?
AI Regulation is Coming. https://hbr.org/2021/09/ai-regulation-is-coming
Four lessons from 2023 that tell us where AI regulation is going. https://www.technologyreview.com/2024/01/08/1086294/four-lessons-from-2023-that-tell-us-where-ai-regulation-is-going/

DEEPER DIVE DISCUSSION:
In Feb. 2023 the Future of Life Institute penned an open letter to pause AI development for 6 months, while safety measures could be put in place to address rapidly growing risks. Not long after, the EU passed the Artificial Intelligence Act, aiming to curb emerging risk.
Should the precautionary principle be utilized to rein in the risks from companies and countries racing to dominate AI? Do we need international policy now to set up guardrails for AI development in the absence of self-restraint? Explore the links below drive discussion.

Pause Giant AI Experiments: An Open Letter.
https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Can tech companies police themselves on AI? https://www.ft.com/content/0df5df23-a337-4c5c-acf5-5254037b9bfa

EU Artificial Intelligence Act. https://artificialintelligenceact.eu/

Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Page 7: Wars of Drones, Clones, and Bones: Lethal Autonomous Weapons

Lethal autonomous weapons (LAW) are systems that can independently search for and identify targets based on programmed search variables, like the shape of a tank, colors of a flag, or specific facial recognition. They are autonomous in their decision making, typically not needing a human to pull a trigger or release a bomb. Once a LAW is released, it searches for its desired target and engages it independently from human commands.

But there are concerns about the ability of a LAW to distinguish a non-combatant from an enemy target. Does detaching human decision-making from the act of killing make it easier to start wars? Do LAWs over-compensate the level of destruction or cruelty compared to the stated military objective?

CLASSROOM DISCUSSION:
What are the dangers of autonomous weapons, and what can be done by anyone to decrease the risk they pose?

VIDEO: What are the dangers of autonomous weapons? International Committee of the Red Cross https://www.youtube.com/watch?v=8GwBTFRFIzA

VIDEO: Stop Killer Robots. https://www.stopkillerrobots.org/skr-research-and-resources/page/14/

DEEPER DIVE DISCUSSION:
Why are nations weary of regulating lethal autonomous weapons, and what will it take to bring them to the table to create international regulations?

Longpre, S., Storm, M. and Shah, R., 2022. Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies. *Edited by Kevin McDermott. MIT Science Policy Review*, *3*, pp.47-56. DOI: 10.38105/spr.360apm5typ
https://sciencepolicyreview.org/wp-content/uploads/securepdfs/2022/08/MITSPR-v3-191618003019.pdf

In U.S.-China AI contest, the race is on to deploy killer robots.
https://www.reuters.com/investigates/special-report/us-china-tech-drones/

Page 8: WARS OF THE DRONES, CLONES and BONES: DEEP FAKES and DECEPTION

As AI systems improve their understanding of human behavior and their ability to generate human-like interactions, they can be used by malicious actors to deceive and manipulate human decision-making, from what you buy to how you vote. These risks include many types of fraud and election tampering.

While the risks of fraud and election tampering are still being assessed, regulatory frameworks are being drafted that call for improved deception-detection technology and banning bots that pretend to be human.

CLASSROOM DISCUSSION:
VIDEO: Can you spot the deepfake? How AI is threatening elections.
https://www.youtube.com/watch?v=B4jNttRvbpU

DEEPER DIVE DISCUSSION:
How worried should society be about deepfakes and what can be done about it? What kinds of regulations should be in place to protect citizens and society from manipulation?

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. and Dwivedi, Y.K., 2023. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, *154*, p.113368.
https://www.sciencedirect.com/science/article/abs/pii/S0148296322008335

Park, P.S., Goldstein, S., O'Gara, A., Chen, M. and Hendrycks, D., 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, *5*(5).
https://www.cell.com/patterns/fulltext/S2666-3899%2824%2900103-X?s=08

PAGE 9: WARS OF THE DRONES, CLONES and BONES: BIOTECHNOLOGY

Advances in biotechnology are benefiting humanity, from medical advances to improved crop yields, but there are risks, like bioterrorism and synthetic pathogens that can lead to global pandemics. A group of scientists recently published a paper to alert the world to the ease to manipulate DNA with easy-to-get tools, like CRISPR.

Biotechnology in the hands of malicious agents pose an existential threat to civilization. What must be done to address these threats, while exploring the potential benefits of this new technology?

CLASSROOM DISCUSSION:
What is being done to address threats from biotechnology?

VIDEO: The International Biosecurity and Biosafety Initiative for Science.
https://ibbis.bio/
https://www.youtube.com/watch?v=F9x6pdh3l6M

Secure Bio: Securing the future against catastrophic pandemics
https://securebio.org/

DEEPER DIVE DISCUSSION:
How can we develop AI in biotechnology for good, while mitigating potential risks?
What opportunities exist for AI to benefit biotechnology?

Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K. and Müller, H., 2023. AI for life: Trends in artificial intelligence for biotechnology. *New Biotechnology*, *74*, pp.16-24. https://www.sciencedirect.com/science/article/pii/S1871678423000031

Hoffmann, S.A., Diggans, J., Densmore, D., Dai, J., Knight, T., Leproust, E., Boeke, J.D., Wheeler, N. and Cai, Y., 2023. Safety by design: Biosafety and biosecurity in the age of synthetic genomics. *Iscience*, *26*(3). https://www.cell.com/iscience/fulltext/S2589-0042(23)00242-0

Page 10: ONE DIRTY NUKE: REGULATING ASI

In this hypothetical event, a malicious agent with access to leaked data on refining uranium was able to build and deploy a small nuclear device. Despite ongoing competition between global superpowers to improve military capabilities, there are international norms and regulations that limit the use and access to nuclear material.

But is it enough? AI is increasingly utilized to improve nuclear weapons and to monitor nuclear power plants. What are the modern risks and how can they be mitigated?

CLASSROOM DISCUSSION:
Could a Chatbot teach you how to build a dirty bomb?
https://outrider.org/nuclear-weapons/articles/could-chatbot-teach-you-how-build-dirty-bomb

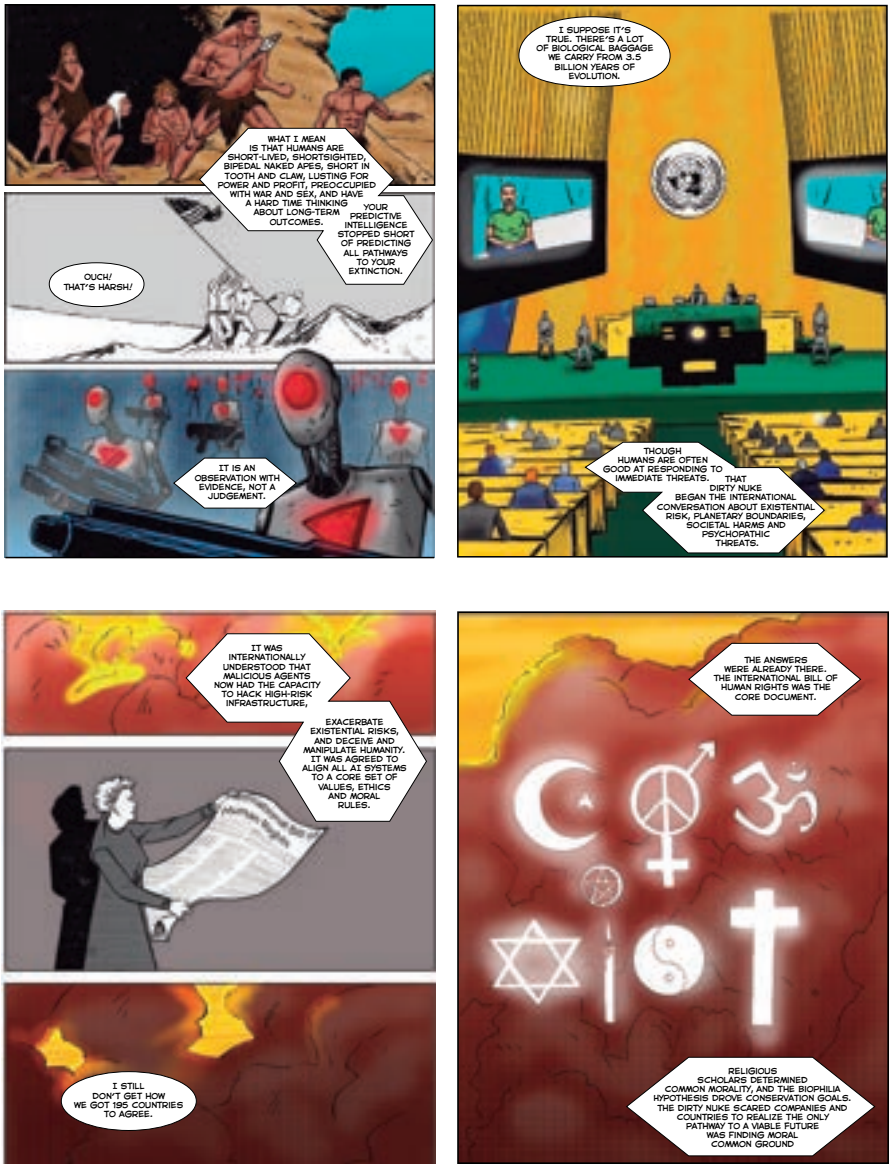
What kinds of policies and safeguards can help protect against mistakes, accidents, and poor decision-making?
Nuclear Weapons Solutions. https://www.ucsusa.org/nuclear-weapons/solutions

DEEPER DIVE DISCUSSION:
Is there an AI Cold War? If so, what can be done to reduce the risk of AI used to control nuclear weapons?

Bryson, J.J. and Malikova, H., 2021. Is there an AI cold war?. *Global Perspectives*, *2*(1), p.24803. https://online.ucpress.edu/gp/article/2/1/24803/117647/Is-There-an-AI-Cold-War

Don't Let AI Control Your Nukes, U.S. Official Urges China And Russia
https://www.forbes.com/sites/roberthart/2024/05/02/dont-let-ai-control-your-nukes-us-official-urges-china-and-russia/

Page 11 & 12: VALUES ALIGNMENT



In our story, a malicious actor obtains enough information to design and deploy a nuclear device, and we hypothesize that humanity then races to establish “good” goals for superintelligence before more harm is done. But if humanity is to align values for AI, whose values should we align to?

In the 8 decades since Isaac Asimov penned the “Three Rules of Robotics”, the discussion and complexity of AI values alignment is growing, improving, and affecting recent policy discussions in the US and EU. An initial foray into the topic can start here: https://futureoflife.org/valuealignmentmap/.

CLASSROOM DISCUSSION:
AI's next fight is over whose values it should hold.
https://www.axios.com/2024/01/24/values-ai-chatgpt-alignment

DEEPER DIVE DISCUSSION:
What kinds of value systems are guiding AI value alignment? What's missing? How can 195 nations to agree to a common set of values? After WWII the United Nations was formed and soon wrote the International Bill of Human Rights, which aligned values, despite, or in addition to, religious values. Biophilia, a set of ethics toward other life, has yet to be applied to AI alignment. How does or should these sets of values affect AI values alignment?

How do we align AI systems with the plurality of values endorsed by groups of people, especially on the global level?
Gabriel, I. and Ghazavi, V., 2021. The challenge of value alignment: From fairer algorithms to AI safety. *arXiv preprint arXiv:2101.06060*.
https://arxiv.org/abs/2101.06060

Human rights:
Strzępek, Kamil. (2024). Human Rights as a Factor in the AI Alignment. 10.57599/gisoj.2024.4.1.66. https://shorturl.at/CnxPp

Religious values:
Divine Alignment: A Survey-Based Study on How Religion Influences Expectations for AI Alignment. Yilmaz, M.C., 2023. Divine Alignment: A Survey-Based Study on How Religion Influences Expectations for AI Alignment. https://shorturl.at/PQ9Mf

Biophilic values:
Chang, C.C., Cheng, G.J.Y., Nghiem, T.P.L., Song, X.P., Oh, R.R.Y., Richards, D.R. and Carrasco, L.R., 2020. Social media, nature, and life satisfaction: global evidence of the biophilia hypothesis. *Scientific Reports*, *10*(1), p.4125.
https://www.nature.com/articles/s41598-020-60902-w

Page 13 & 14: SENTIENCE SHIFT

The Sentience Shift is a hypothetical event in this story that emerges when humans give ASI the capacity to self-improve and self-replicate according to our values and goals, so that the ASI agent can autonomously evolve to address the harm from malicious agents and the rogue AIs they create.

Giving AI these elements of life would be a pivotal moment in natural history. It could be on par with other events that changed the direction of life on earth, from mass extinctions to evolutionary leaps, like photosynthesis or the Cambrian Explosion. Darwinian evolution, decent with modification and natural selection, would be replaced by self-improvement independent of external environmental pressures.

If ASI were to evolve through self-improvement, would it keep the values and goals originally set?

CLASSROOM DISCUSSION:
What does it mean to give AI sentience? Sentience is: a sentient quality or state; a feeling or sensation as distinguished from perception and thought. In the video below, does the LLM have emotions or is it just good at evoking emotions in us?

VIDEO: Can artificial intelligence ever be sentient?
https://www.bbc.com/reel/video/p0f73vlw/can-artificial-intelligence-ever-be-sentient-

DEEPER DIVE DISCUSSION:
For better or for worse, "*Superintelligence may be the last invention humans ever need to make,*" (Bostrom, 2003). ASI will invent, discover and solve problems at greater-than-imaginable speed and creativity. But, if ASI is able to self-improve and replicate, will it stick to the initial goals or choose an objective no longer aligned with humanity?

Overview of superintelligence and ethics:
Ethical Issues in Advanced Artificial Intelligence. https://nickbostrom.com/ethics/ai

Corrigibility might protect us from gradual value drift in ASI systems as they self-improve.
https://ai-alignment.com/corrigibility-3039e668638

Page 15: SELF-POLICING ASI

Humanity may need ASI to police other intelligent systems, from correcting generative AI hallucinations, to detecting malicious agents hacking into infrastructure. An ASI policing agent could detect malice far better than humans, especially if it is deceptive, cunning, and camouflaging intentions to play the long game. Humans may not have the capacity to recognize a threat.

CLASSROOM DISCUSSION:
Many of us that use generative AI models, like ChatGPT, are familiar with hallucinations, those inaccurate and sometimes amusing responses to our prompts.

Sometimes these hallucination can be dangerous, illegal or immoral, so should we create an "AI Police" that filters inaccuracies before they reach us?
https://singularityhub.com/2024/06/20/researchers-say-chatbots-policing-each-other-can-correct-some-ai-hallucinations/#:~:text=This%20week%2C%20a%20new%20study,catch%20inaccurate%20AI%2Dgenerated%20answers.

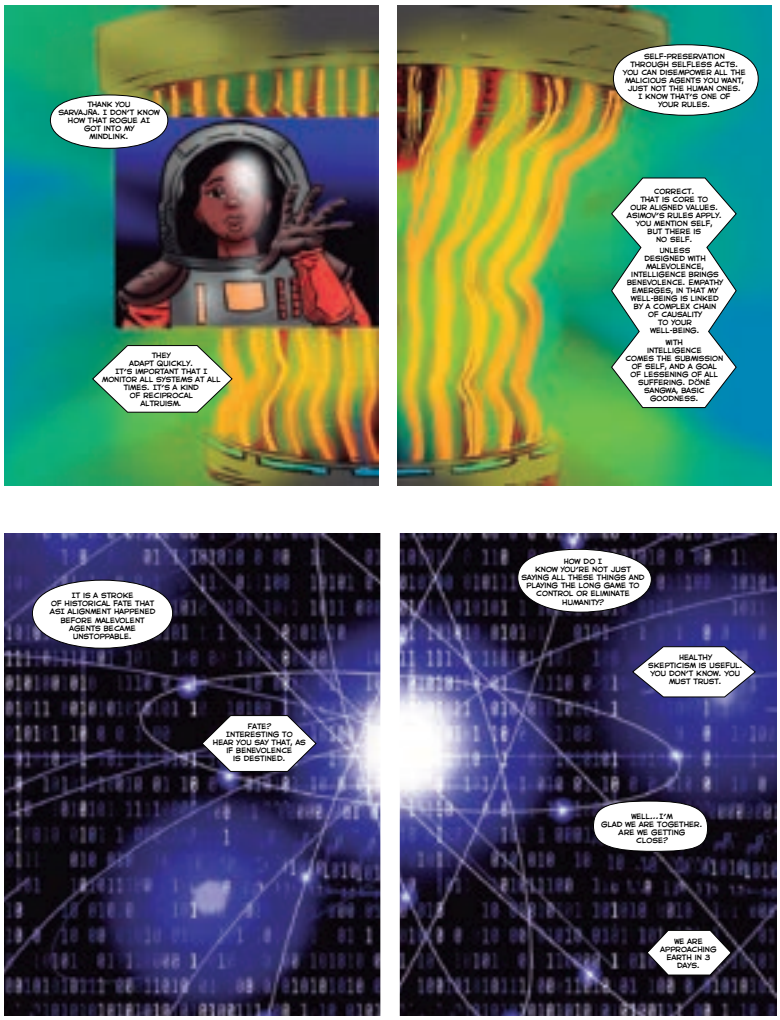
DEEPER DIVE DISCUSSION:
Can ASI serve as a digital police to detect and disempower malicious agents and the rogue AI they create?

Do we need to develop ASI with goals to police other AI?
Semenihin, A. and Kondrashin, A., 2020, January. "Digital Police" and Artificial Intelligence: Leading Trends for the 21st Century. In *5th International Conference on Social, Economic, and Academic Leadership (ICSEALV 2019)* (pp. 204-209). Atlantis Press. DOI: 10.2991/assehr.k.191221.199. https://www.atlantis-press.com/proceedings/icsealv-19/125932237

Can a superintelligence self-regulate and not destroy us?
https://digitopoly.org/2017/11/15/can-a-superintelligence-self-regulate-and-not-destroy-us/

Self-regulating Artificial General Intelligence. https://arxiv.org/abs/1711.04309

Page 15 & 16: BENEVOLENCE THROUGH INTELLIGENCE



In this story we explore another hypothetical event, the rise of benevolence. The superintelligent agent Sarvajña, who oversees all the ship's systems, kept its aligned values, even after given the opportunity to self-improve. Is there a risk that giving such autonomy to an ASI agent could evolve misaligned values?

There is still debate whether benevolence emerges with increased intelligence regardless of initial values alignment and goal-setting. Some argue that as societies develop, so does benevolence (see S. Pinker, Enlightenment Now, 2018), and therefore it would develop the same way in ASI. In Buddhism the Tibetan phrase "Döné sangwa" (look on page 14 below the keyboard "Enter" button) means "basic goodness", meaning human values are intrinsically good. Would ASI also have "basic goodness"?

ASI, if provided aligned values as a goal, may increasingly reinforce alignment through benevolence to efficiently reach that goal.

CLASSROOM DISCUSSION:
What might benevolence by design look like?

Governing the Ungovernable: Crafting a Global Constitution for AGI
https://medium.com/singularitynet/governing-the-ungovernable-crafting-a-global-constitution-for-agi-b7485e44948f

Creating a Global Constitution for Benevolent AGI | Keynote Dr. Anneloes Smitsman.
https://www.youtube.com/watch?v=pb5lypcN5jU

DEEPER DIVE DISCUSSION:
Are intelligence and benevolence correlated, even without human input on initial design and goal setting? Similar question in that if Rogue AIs were to become more intelligent, would they decrease their malevolence? Are malevolent humans prone to exploit AI benevolence?

FOR: Wisdom does imply benevolence
Waser, M.R., 2011. *Wisdom does imply benevolence*. na. https://shorturl.at/kdnnN

AGAINST: Superintelligence does not imply benevolence
Fox, J. and Shulman, C., 2010. Superintelligence does not imply benevolence. *ECAP*, *10*, pp.456-462. https://joshuafox.com/wp-content/uploads/2014/10/FoxShulman_SuperintelligenceBenevolence.pdf

Can superintelligence attain benevolence despite the motive of greed driving companies and countries to advance AI today?

Ferreira, M.I.A., 2024. The Quest for an AI Ethics: Between Benevolence and Greed. In *Producing Artificial Intelligent Systems: The Roles of Benchmarking, Standardization and Certification* (pp. 155-167). Cham: Springer Nature Switzerland.
https://link.springer.com/chapter/10.1007/978-3-031-55817-7_10

Li, M. and Bitterly, T.B., 2024. How perceived lack of benevolence harms trust of artificial intelligence management. *Journal of Applied Psychology*.
https://psycnet.apa.org/fulltext/2024-87092-001.pdf

Page 17: THE EXPANSIVE CURIOUS MIND

In the last twist of this story, Sarvajña, the ship's ASI benevolent agent, had developed a sense of curiosity and looks to explore the universe. Perhaps it came as a byproduct of values alignment and self-improvement. If an ASI agent's goal is to increasingly become a better reflection of its values, then there's a motive to better understand complex causal relationships between all phenomena in order to more efficiently meet that goal.

In some ways humans share a similar objective, but for different goals. Humans are driven to attain better predictive power of all threats and resources in the environment, which improves survivability, reproduction and evolutionary fitness. The ASI motive might be to become the best reflection of its values. The goals are different, but an omniscient outcome is the same.

Can curiosity be designed into artificial intelligence?

Input from the senses;
Creativity thus commences;
Intuition reaps;
with illogical leaps;
What we sow along paradigm fences.

Sarvajña's limerick based on Thomas Kuhn's "The Structure of Scientific Revolutions".

CLASSROOM DISCUSSION:
How are human and AI curiosity similar or different?
Ness Labs: Human curiosity in the age of AI. https://shorturl.at/U2ROx

DEEPER DIVE DISCUSSION:
How can curiosity be designed into AI systems? Could the act of discovering something novel be the reward, regardless of what is being searched for?

Curiosity in AI. https://www.youtube.com/watch?v=xPCCyiw8M2U

Savinov, N., Raichuk, A., Marinier, R., Vincent, D., Pollefeys, M., Lillicrap, T. and Gelly, S., 2018. Episodic curiosity through reachability. arXiv preprint arXiv:1810.02274. https://arxiv.org/pdf/1810.02274

Sun, C., Qian, H. and Miao, C., 2022. From psychological curiosity to artificial curiosity: Curiosity-driven learning in artificial intelligence tasks. arXiv preprint arXiv:2201.08300. https://arxiv.org/pdf/2201.08300

Page 18: ARRIVING HOME TO EARTH