Sentiment Analysis of Reddit Posts on AI in Education

# Brief Summary of Findings

This analysis of Reddit posts reveals a sharply polarized conversation surrounding AI in education. Negative sentiment is predominant and deeply personal, with educators expressing concerns about academic integrity, the erosion of student critical thinking, and a sense of 'moral injury' from increased workload and institutional pressures. In contrast, positive sentiment is more philosophical and forward-looking, focusing on the potential for AI to engage in logical, ethical reasoning. Neutral posts are typically factual, news-based announcements, indicating that emotional responses are tied more to the practical application of AI rather than its mere existence.

# Top 3 Most Positive 🌟 Posts

## A New Way to Educate

**Subreddit:** r/education  
**Sentiment Score:** 1.00  
**URL:** https://www.reddit.com/r/education/comments/1cvdu41/a\_new\_way\_to\_educate/

Exploration of an Innovative K-12 School Curriculum and Pedagogical Approaches  
  
  
Before going over the curriculum we must first discuss how to best teach. All of the following methods outlined below are what would I would suggest be use to teach the students trying to use just one or even just two of these would not be enough and would compromise the students learning and education  
  
  
Teaching Methods   
  
Project-Based Learning (PBL): In PBL, students work on a project over an extended period, which could be a week or a month, to respond to a complex question, problem, or challenge. The projects are usually multi-disciplinary and require students to apply what they've learned in a practical manner. This allows them to see the immediate applicability of their learning.  
  
Inquiry-Based Learning: This is a form of active learning that starts by posing questions, problems, or scenarios—rather than simply presenting established facts or portraying a smooth path to knowledge. Students are involved in the construction of their learning. They engage with the material, participate in the class, and collaborate with each other.  
  
Gamification: Incorporating elements of game design in education can make learning fun and engaging. This can involve point systems, leaderboards, badges, or other game mechanics.  
  
Experiential Learning: This method involves learning by doing and reflecting on the experience. It can include internships, study abroad programs, field trips, laboratory experiments, or any other hands-on learning experiences.  
  
Flipped Classroom: In a flipped classroom, students review lecture materials at home and do their 'homework' in class, where they can ask for help as they practice new skills and apply new knowledge. This allows teachers to spend class time helping students apply what they've learned and coaching them as they work through challenges.  
  
Cross-Disciplinary Projects: By integrating different disciplines into a single project, you can make the learning experience more holistic and interconnected, much like how the knowledge of different magical disciplines would combine in a fantasy setting.  
  
Competency-Based Learning: In this educational model, students advance upon mastering a skill or a competency. This encourages active utilization of knowledge and immediate feedback, similar to how a magic student might advance only after successfully casting a particular spell.  
  
Now that we have given a basic outline of the teaching styles we can go over the curriculum for K-12 the idea would be to Have an A/B day schedule and some classes would meet less frequently because they don't take much time to cover everything, all of this will be done in order to create well rounded students, people and citizens. They are not only creative in nature, but leaders in their own right as well as capable of doing whatever they desire and succeeding wildly  
  
  
Core Curriculum Classes:   
  
STEAM (Science, Technology, Engineering, Art, Mathematics): must Ensure these subjects are covered both within the integrated curriculum (like coding in math, cooking in science, history in art, etc.) and as standalone classes to develop depth of knowledge.  
  
  
Memory Techniques & Knowledge Management Techniques: Using an Integrated code based system with AI tools to help teacher track progress and provide more targeted assistance as well as help students with how to effectively organize and manage knowledge, covering basic note-taking, the PARA/CODE system. Using AI to provide semi interactive sessions that not only explain how memory works but also actively encourage the practice of using the note taking method & memory techniques  
  
  
Project-Based Learning: Encourage practical application of knowledge through project-based learning.  
Gamification & Experiential Learning: Use these techniques to make learning more engaging and fun.  
  
  
Flipped Classroom & Inquiry-Based Learning: Encourage independent learning and critical thinking through these teaching methods.  
  
  
  
Elementary School:  
Coding & Digital Literacy: Introduce basic coding principles using visual coding platforms. Begin teaching about online safety and basic cybersecurity.  
Financial Literacy: Teach basic concepts like the value of money, saving, and spending. Potential to introduce the use of real currency and creating student based economy  
Community Service: Arrange class-based community projects and encourage involvement in community service outside of school.  
Gardening & Cooking: Teach students about plants, nutrition, and basic cooking skills through a school garden.  
Literacy & Reading: Develop a reading program that exposes students to a variety of genres. Writing can begin with simple sentences.  
History: Teach history from a holistic and critical perspective, exploring different cultures and perspectives.  
Basic Medicine & First Aid: Introduce simple health, hygiene, and basic first aid skills.  
Physical Education: Encourage a love for physical activity through a variety of engaging games.  
Emotional Regulation & Healthy Relationships with Technology: Incorporate social-emotional learning and healthy technology use.  
Leadership: Begin fostering leadership qualities through group activities and responsibility sharing.  
Self-Defense: Introduce basic safety rules and personal boundaries.  
Spanish: Introduce basic Spanish vocabulary and phrases, along with exposure to the culture of Spanish-speaking countries. Songs, games, and interactive activities can be used to make learning enjoyable.  
  
  
Middle School:  
Coding, Digital Literacy & Practical Engineering: Continue coding education and introduce robotics and basic electronics.  
Financial Literacy: Start teaching about budgeting, banking, and simple concepts of earning. Potential to introduce the use of real currency and creating student based economy  
Community Service: Encourage students to plan and lead community service projects, either in groups or individually.  
Gardening & Cooking: Progress in gardening and cooking skills, introducing sustainability issues.  
Literacy, Reading, & Writing Skills: Increase complexity of reading and writing assignments.  
History: Provide more in-depth history education using primary sources and interpretations.  
Basic Medicine & First Aid: Offer a more detailed course on first aid and health.  
Physical Education: Introduce a range of physical activities, sports, and body awareness topics.  
Emotional Regulation & Healthy Relationships with Technology: Develop emotional intelligence skills, mindfulness practices, and education about responsible technology use.  
Leadership: Teach various leadership styles and emphasize group projects requiring delegation and decision-making.  
Self-Defense: Continue with more practical self-defense techniques.  
Study of Government: Begin a foundational study of the local and national government. Teach students about the branches of government, their roles, and how laws are made.  
Spanish: Continue to build on vocabulary and grammar learned in elementary school. Introduce simple written exercises and encourage basic conversation in Spanish.  
  
High School:  
Coding, Digital Literacy & Practical Engineering: Offer advanced coding and practical engineering classes, including topics like 3D modeling and advanced electronics.Teach more advanced cybersecurity concepts and ethics of digital communication  
Gardening & Cooking: As elective courses, delve into advanced topics like food science or agricultural technology.  
Literacy, Reading, & Writing Skills: Offer a variety of literature courses, creative writing classes, and research-based writing.  
History: Teach history as a dynamic and interpretive subject, encouraging critical thinking.  
Basic Medicine & First Aid: Include more advanced first aid, mental health awareness, and basic human anatomy and physiology.  
Physical Education: Offer a range of athletic options, and include education about exercise science and long-term health benefits.  
Emotional Regulation & Healthy Relationships with Technology: Provide resources for emotional regulation, advanced mindfulness techniques, and in-depth discussions about technology's role in society.  
Leadership: Delve deeper into conflict resolution, strategic planning, and ethical leadership, often through real-world applications.  
Financial Literacy: Start teaching about budgeting, banking, and simple concepts of earning. Potential to encourage student to start their own business or get a job and have students buy things from each other   
Self-Defense/Mixed Martial Arts: For interested students, offer elective classes in mixed martial arts, fostering physical and mental skills.  
Study of Government: Expand on knowledge from middle school and introduce international government systems. Discuss the concepts of democracy, socialism, and other forms of government. Involve students in mock debates and simulations, like Model United Nations or Mock Parliament.  
Study of Politics: Begin a course on political science, covering key political ideologies, parties, and political processes. Discuss current events and involve students in debates and discussions to encourage critical thinking.  
Creating Change: Introduce a course on social activism and creating change. This can involve studying historical movements for change, understanding how to effect change within a legal framework, learning about peaceful protest, and planning and implementing a small-scale change project within the school or community.  
Spanish: Continue to deepen knowledge of Spanish. Encourage advanced conversation and written exercises. Students could read Spanish literature or news and discuss in class, fostering language skills and cultural understanding.  
  
This kind of curriculum would be nearly the best, being interdisciplinary, hands-on, and centered around the interests and needs of the students. It would aim to not only equip students with the knowledge and skills they need to succeed in the world, but also ignite their passion for learning and encourage them to continue learning throughout their lives.  
  
But before we’re done one last thing must be covered. How to assess a student's growth because science shows that paper tests are not suited for the task. There are many innovative ways to assess student understanding and skills without relying solely on traditional exams. The methods that could be used include  
Assessment Methods   
  
Portfolios: Students could compile portfolios of their work, which could include code they've written, projects they've completed, or essays they've written. A portfolio allows students to demonstrate their learning process, their progress over time, and their ability to apply what they've learned in different contexts.  
  
Presentations: Students can demonstrate their understanding of a topic by presenting on it. This could involve presenting a project they've completed, explaining a concept to the class, or debating a topic with classmates.  
  
Peer and Self-Assessment: Students can learn a lot from assessing each other's work or their own work. This can help them develop a better understanding of the assessment criteria and improve their ability to critically evaluate work.  
  
Performance Assessment: In subjects like self-defense, physical education, cooking, or gardening, students could be assessed based on their performance. This could involve demonstrating a technique, completing a task, or participating in a game or competition.  
  
  
Reflective Journals: Students could maintain journals where they reflect on what they've learned, how they've applied it, and what they still want to understand better. This can give teachers insight into a student's thought process and their understanding of the subject.  
  
  
Project-Based Assessment: Students can be assessed on the projects they complete, whether individually or in groups. This allows students to demonstrate a range of skills, including knowledge of the subject, problem-solving, creativity, and teamwork.  
  
  
Community Service Assessment: In addition to the other assessments, teachers can assess students' community service involvement, their planning and leadership skills, as well as their reflections on their experiences.  
   
  
The emphasis of Knowledge Management Techniques and Memory Techniques in core classes as a standalone session every day would ideally give students a break from traditional instruction and allow them to process and manage their learnings. This can be in the form of group discussions, independent reflection time, or guided activities for planning and organizing their work.

## Notes from the NYU AI Ethics conference

**Subreddit:** r/AIethics  
**Sentiment Score:** 1.00  
**URL:** https://www.reddit.com/r/AIethics/comments/57vno0/notes\_from\_the\_nyu\_ai\_ethics\_conference/

This weekend I attended the [Ethics of Artificial Intelligence conference](https://wp.nyu.edu/consciousness/ethics-of-artificial-intelligence/) at NYU. There were a ton of high-profile and interesting people there from philosophy (David Chalmers, Peter Railton, Nick Bostrom, Thomas Nagel, Paul Boghossian, Frances Kamm, Wendell Wallach) and science (Yann LeCun, Stuart Russell, Stephen Wolfram, Max Tegmark, Francesca Rossi) as well as Eliezer Yudkowsky.   
  
There were two fairly long days of talks and panels. David Chalmers (famous for his philosophy of mind and consciousness) did not officially speak but acted as chair for the event. He outlined the philosophy of the conference, which was to discuss both short and long term issues in AI ethics without worrying about either detracting from the other. He was, as usual, extremely awesome.  
  
Here is a summary of the event with the most interesting points made by the speakers.  
  
\*\*Day One\*\*  
  
The first block of talks on Friday was an overview of general issues related to artificial intelligence. Nick Bostrom, author of \*Superintelligence\* and head of the Future of Humanity Institute, started with something of a barrage of all the general ideas and things he's come up with. He floated the idea that perhaps we shouldn't program AI systems to be maximally moral, for we don't know what the true morality looks like, and what if it turns out that such a directive would lead to humans being punished, or something else that was pathological or downright weird? He also described three principles for how we should treat AIs: substrate nondiscrimination (moral status does not depend on the kind of hardware/wetware you run on), ontogeny nondiscrimination (moral status does not depend on how you were created), and subjective time (moral value exists relative to subjectively experienced time rather than objective time, so if a mind ran at a fast clock speed its life would be more important, all other things being equal).  
  
He pointed out that AI moral status could arise before they reach there is any such thing as human level AI - just like animals have moral status despite being much simpler than humans. He mentioned the possibility of a Malthusian catastrophe from unlimited digital reproduction as well as the possibility for vote manipulation through agent duplication, and how we'll need to prevent these two things.  
  
He voiced support for meta level decisionmaking - a ['moral parliament'](http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html) where we imagine moral theories sending 'delegates' to compromise over contentious issues. Such a system could also accommodate other values and interests besides moral theories.  
  
He answered the question of "what is humanity most likely to fail at?" with a qualified choice of 'mind crime' committed against advanced AIs. Humans already have difficulty with empathy towards animals when they exist on farms or in the wild, but AI would not necessarily have the basic biological features which incline us to be empathetic at all towards animals. Some robots attract empathetic attention from humans, but many invisible automated processes are much harder for people to feel empathetic towards.  
  
Virginia Dignum was next; she is at the Delft University of Technology and spoke about mechanisms for automated processes to make decisions. She specified four methods of decisionmaking based on whether decisions are taken deliberately or imposed upon a system and whether the decisions are made internally or externally. The two former features lead to algorithmic decisionmaking in machines; the latter two lead to imposed decisions predetermined by regulatory institutions. Deliberated external decisionmaking means there is a 'human in the loop' and internal imposed decisionmaking is essentially randomness.  
  
Yann LeCun concluded this section with a pretty fantastic overview of deep learning methods and the limitations which stand in the way of progress in machine intelligence. He pointed out that reinforcement learning is a rare and narrow slice of the field today and that the greatest obstacles for machines include common sense judgements and abstraction. The biggest current problem for AI is unsupervised learning, which is having machines that can learn to classify things on their own without being given clearly labelled data from humans. He showcased some of the (very cool) features of adversarial learning which are being used to tackle this.  
  
He expressed support for the orthogonality thesis, namely the idea that intelligence and morality are 'orthogonal' - just because an agent is very smart doesn't mean that it's necessarily moral. He believes we should build a few basic drives into AIs: do not hurt humans, interact with humans, crave positive feedback from trusted human trainers. He also described a couple of reasons for why he is not concerned about uncontrolled advanced artificial intelligence. One was that he is confident that objective functions can be specified in such a way as to make machines indifferent to being switched off, and the other is that a narrow-AI focused on eliminating an unfriendly general-AI would 'win' due to its specialization.  
  
In Q&A, Stuart Russell objected to LeCun's confidence in machines being indifferent to being shut off based on the fact that self-preservation as a goal implicitly falls out of whatever other goals a machine has. Paul Boghossian objected to the 'behaviorist' nature of the speakers' points of view, saying that they were exempting consciousness from its proper role in these discussions. One person asked whether we should let AIs take charge of everything and supersede humanity - Bostrom pointed out that the space of possible futures is "an enormous Petri dish" which we don't understand; an AI future could materialize as a planet sized supercomputer with no moral status, and we will need to learn how to engineer friendly advanced AI systems no matter what the plan is.   
  
The rest of the Friday talks were devoted to near-future issues with specific AI systems. Peter Asaro started with an overview of his organization, the 'Campaign to Stop Killer Robots'. He stated that targeting and killing should remain human-controlled actions. While he acknowledged that automated weaponry could result in fewer casualties on the battlefield, he believed that it was too narrow a view of the consequences. He said that it's not straightforward to translate complicated battlefield morality questions for machines to understand, and is worried about unintended initiation and escalation of conflicts through automated systems, arms races, and threats to humanitarian law. He also believes that people should only be killed with 'dignity' and that doing it with a robot robs people of this. Therefore, he called for a clear and strong norm against automated weapons.  
  
Kate Devlin of the University of London gave a brief overview of the ethics of artificial sexuality. Looking at the history of sexualized robots featured in fictional media, she noted that almost all of them are female. Today there is a "Campaign Against Sex Robots" which is based on the idea that sexual robots would lead to the objectification of women. Devlin does not agree as she thinks it is too early to ban the technology and that we should explore it before thinking about banning it, especially since it does not really harm anyone. Instead she wants us to think about how to develop it correctly. There are many potential uses for these types of robots ranging all the way to the therapeutic; many of the rudimentary ones being sold today are bought by people who are incapable of forming ordinary relationships for various reasons. VR is being used in arousal tests to gauge the efficacy of treatments against pedophilia.  
  
She noted that gender issues have arisen in technology already; the history of gendered technology includes pacemakers originally designed only for men and phones too large for women's pockets. We should get into AI now to make sure that it is not designed in problematic ways.  
  
She mentioned privacy concerns, as the manufactures of the female stimulator WeVibe have already been sued over concerns that they were not properly informing customers of their collection of data from the devices. She wondered if we will ever get to a stage where a robot might have some knowledge of its role and refuse to give consent to its use, and if transmission/duplication of data and code between machines could serve as some form of digital sexual reproduction.  
  
Vasant Dhar of NYU spoke next about data and privacy in the era of autonomous vehicles. He said that our legal and financial liability institutions are based on outdated notions of data and that they fail to address liability and crime. However, the tools we have now even in ordinary cars for recording data can be used to improve insurance and judicial systems. He proposed black boxes for cars that would contain all relevant data to determine fault in the event of accidents, and said that customers should have the choice to share their driving data with insurance companies to get lower premiums.   
  
Dhar reiterated the importance of improving vehicle safety through autonomous driving; each percentage point reduction in vehicle accidents equates to 400 deaths and 40,000 injuries avoided every year.  
  
Adam Kolber followed up with a discussion of whether "the code is the law", based on the case study of [The DAO](https://en.wikipedia.org/wiki/The\_DAO\_%28organization%29) which was an automated capital fund which was subjected to a $50 million loss through exploitation. The answer apparently is that the code should not be the law, even though many people seemed to accept that it was.  
  
Steve Wolfram of WolframAlpha and Mathematica fame discussed the issues of computer languages and goal specification. He said that his life work has essentially been about trying to find ways for humans to specify their goals to machines, and that this can work for ethics as well as for math. He doesn't think that any single moral theory is likely to work for guiding artificial intelligence, apparently because of Godel's theorem and the incompleteness of computational languages.  
  
Francesca Rossi of IBM argued that for AIs and humans to interact very productively we will have to embed them in environments, so that rather than picking up a tool like a laptop or a phone, we are interacting with artificial systems all around is in our rooms and spaces. Humans will be recognized by their environments and our needs and wants will be inferred or asked about. AI embedded in environments can have memories about humans to better serve their interests. Most of all, we will need to establish trust between humans and AIs.  
  
Peter Railton, philosopher at the University of Michigan, attacked the subjects of orthogonality and value learning. He said that we can't simply tell AIs to do what we want because our wants and values require critical assessment. He said that the orthogonality thesis might be right, but as we increasingly interact with systems and allow them to participate in our own lives and decisionmaking, the question of what it would take for them to be intelligent might involve certain features relevant to morality.  
  
He stated that AIs should be thought of as social creatures; as a simple model, self regulation in a Hobbesian social contract leads to constraints and respect derived from self preservation. A society of intelligent cooperators can resist aggression and malice, and being moral is more efficient for a community than being cunning. From these principles we have a recipe for building proto-moral agents.  
  
He discussed the 'moral point of view' required for many strong ethical theories such as Kantian ethics and consequentialism: it requires agents to have a hierarchical, non-perspectival, modal/planning-oriented, and consistent view of the world which assigns intrinsic moral weight to things. He described how all these features are also part of the process of becoming generally intelligent in the first place, implying that general social intelligence ensures the necessary information required for moral decisionmaking. In the path towards functional moral agents, we will have to build agents which can represent the goals of others and have them learn how to act in beneficial ways. So if we can build AIs that we can trust, then we are on a good path towards building artificial moral agents.  
  
In the Q&A, Eliezer Yudkowsky objected that in the long run the 'instrumental strategy' is not quite what you want because maximizing people's desires as they are explicitly revealed can lead to bad outcomes, and you have to have a view like coherent extrapolated volition which asks what people would really want. Russell objected that when an agent becomes sufficiently powerful, it has no need to cooperate anymore.  
  
Regina Rini of the NYU Center for Bioethics stated that the approaches to ethics so far described relied too much on the Western post-enlightenment view of ethics, which is a historical aberration, and excluded African, Chinese and other approaches to ethics. Railton stated that his scheme was grounded in basic empathy and not mediated by any higher order moral theory; Wolfram and Rossi said that no one ethical approach will work and AI will have to represent diverse values.  
  
\*\*Day Two\*\*  
  
Saturday was devoted to long term discussion of the future of advanced artificial intelligence. Stuart Russell, professor at UC Berkeley and head of the new Center for Human Compatible Artificial Intelligence, started with a basic overview of the control problem. He described the points made in Steve Omohundro's paper on convergent instrumental drives. He also had some pretty harsh words for the researchers in the AI community which have denied and rejected notions of the control problem without seriously engaging with the relevant literature.  
  
He had three simple ideas which he proposed to constitute the definition of 'provably beneficial' AI: maximizing values for humans is the system's only goal; the robot is initially uncertain about these goals, and the best source of information is human behavior. He referred to inverse reinforcement learning as a technique for machines to learn human preferences, and said that uncertainty provides an incentive for machines to learn, ask questions, and explore cautiously.   
  
His answer to the off-switch problem is to make robots unsure of their objectives, so that they assume that the human will switch the robot off if and only if it has a good reason to, and will therefore be complicit with the action. He said that the wireheading problem can be avoided if you construct the reward signal as information about the reward function rather than as a reward itself; this way, any hijacking of the reward signal makes it useless.  
  
He said that there is a strong economic incentive for value alignment, but humans are irrational, nasty, inconsistent, and weak-willed.  
  
The next speaker was Eliezer Yudkowsky of the Machine Intelligence Research Institute. Chalmers pointed out his role there as well as his side venture in Harry Potter fanfiction.  
  
Yudkowsky started [his talk](https://intelligence.org/nyu-talk/) by pointing out how the Terminator pictures in every media article about the control problem are inappropriate. The real analogy to be used is [Mickey Mouse as the Sorcerer's Apprentice in \*Fantasia\*.](https://www.youtube.com/watch?v=Ait\_Fs6UQhQ)   
  
He said that the first difficulty of AI alignnment is that the utility functions we imagine are too simple, and the second difficulty is that maximizing the probability of achieving a given goal leads to pathological outcomes. He and MIRI are concerned with the nature of the goal of 'maximizing' and how to define goals in a way that avoids the problems of perverse instantiation.   
  
He said that the fears of AI being developed by some terrorist or rogue group were silly, as "ISIS is not developing convolutional neural nets." Instead the most powerful AI is likely to be developed by large groups in government, academia and industry.  
  
He claimed that the four central propositions which support the idea that AI is a very big problem are: the orthogonality thesis, instrumental convergence, capability gain (the speed at which advanced AI can make itself better), and alignment difficulty. He said the first two are logical matters of computer science that people always learn to accept when they reflect upon them, while the latter two are more controversial.  
  
The next talk was from Max Tegmark and Meia Chita-Tegmark. Max is a world-renowned physicist who helps run the Future of Life Institute, and Meia is a psychologist. They explained how physics and psychology provide useful tools for understanding artificial intelligence; physics tells us about computation and the constraints of the universe, and psychology tells us about the nature of well being, ways to debug the mind when reasoning about AI and methods to design psychomorphic AIs. Meia was the only speaker at the conference to discuss unemployment in any detail; she pointed out that retirement has only mixed effects on well being and that happiness comes from financial satisfaction and feelings of respect. She said that studying homemakers, part time workers and early retirees can tell us more about how an automated economy would affect people's well-being.  
  
Max checked off [a list of common myths](http://futureoflife.org/background/aimyths/) regarding advanced AI. Meia said that we should look at the cognitive biases which have led to these misconceptions (such as availability bias leading to people worrying about robots rather than invisible artificial intelligence) and figure out how to avoid similar bugs from inhibiting our thinking in the future.  
  
By the way, Max Tegmark is very cool, he has a sort of old-rocker-dude vibe, and he and Meia are super cute together.  
  
Wendell Wallach of Yale spoke next. He is the man who quite literally wrote [the book](https://www.amazon.com/Moral-Machines-Teaching-Robots-Right/dp/0199737975) on AI ethics. He distinguished top-down approaches of formally specifying AI behaviors from bottom-up approaches of value learning. He said that neither will be sufficient on its own and that both have important roles to play. He is worried that AI engineers will make simplistic assumptions about AI, such as the idea that every decision should be utilitarian or the idea that 'ethics' and 'morality' are icky concepts that can be ignored.  
  
Steve Petersen, a philosopher at the University of Niagara, gave the next talk, based on the draft of a forthcoming paper of his. He aims to push back against the orthogonality thesis and modulate the level of the risk assessment provided by Bostrom. His argument is that designing AI to follow any complex goal will necessarily require it to be able to learn the values of its "teleological ancestors" (the original human designers or the previous iterations of AI before it self-improved or self-modified) and arrive at a state of coherence between goals. As agents replicate, self-modify and merge in the digital world, there can be no fact of the matter about which agents are the same or different; instead there will be an 'agential soup' unified by a common teleological thread originating with the designers. Coherence reasoning leads to impartial reasoning with the goals of other agents.  
  
There were several responses to him in Q&A. Yudkowsky's objection was that reaching coherence requires a meta-preference framework with particular assumptions about the universe and ontology; therefore, for any goal, there are many preference frameworks which could fulfill it, many of which would be perverse. Russell said that just coherence is not enough because you need the systems to give special weight to humans. Max Tegmark said that the problem was the vagueness of humanity's final goals. Chalmers pointed out that the orthogonality thesis still allows for all kinds of correlations between between intelligence and morality, as long as they are not necessary by design. Petersen said that he is arguing for 'attractor basins' in the possibility space of AI minds. Interestingly, he was motivated to start his research by the [Dylan Matthews Vox article](http://www.vox.com/2015/8/10/9124145/effective-altruism-global-ai) on effective altruism where Dylan thought that effective altruists shouldn't be concerned by artificial intelligence. Petersen doesn't think that AI is unimportant and thinks that Bostrom and Yudkowsky's work is valuable, but he wanted to get a more critical assessment of the level of risk when he learned that alternative altruistic projects were at stake.  
  
Matthew Liao of the NYU Center for Bioethics gave an argument for moral status on the basis of capabilities - that an entity is morally valuable to the extent that it has the physical/genetic basis for achieving features of moral relevance. I did not get a chance to ask him if this would imply that a 'seed AI' could be the most morally valuable entity in the world. He did argue against the ideas that level of intelligence or degree of moral agency determine moral status, as we don't normally think that smarter or more benevolent humans are more morally valuable than others.   
  
Liao argued that moral theories are too specific and too high level to be generally implemented in AIs. Instead, AI will need a universal moral grammar in which to specify morality. The holy grail is to develop machines that understand why things are right or wrong.  
  
Eric Schwitzgebel and Mara Garza of UC Riverside argued for basic principles of AI rights. They introduced a very weak "no-relevant-difference" argument: the idea that there are possible AIs which have the same morally relevant features that humans do and therefore there are possible AIs with equal value to humans. They questioned if cheerfully suicidal or hardworking AI is acceptable, and stated a 'self respect principle': that human grade AI should be designed with an appropriate appreciation of its own value.  
  
John Basl and Ronald Sandler of Northeastern University argued for AI research committees to approve or deny research in cases where AI subjects might be harmed. They said it would not be very different from cases like animal testing where we have similar review boards, and sketched out details of how the proposal would work.  
  
Daniel Kahneman, one of the most famous behavioral economists in the world, made something of a surprise appearance in the final panel. He said that we should take intuitions about case studies like the trolley problem seriously, as that is how the public will think about these events, for better or for worse. He said that no matter how AI cars kill people, it will be perceived with horror whenever the first incident happens, and we should prepare for that. Intuitions depend on irrelevant factors and will especially depend on whether AIs are designed to resemble us or not.  
  
Gary Marcus, professor of psychology at NYU, of gave a much needed presentation about the nature of intelligence. The previous talks in this discussion had mostly assumed that intelligence was one-dimensional and simple and that there was some fixed idea of 'human-level' AI which we could eventually reach. Of course this is a ridiculous oversimplification; intelligence is multidimensional and it is more about implementing a combination of various cognitive tools, some of which are already stronger in AIs than in humans. AIs can be better or worse than us in various domains, so we really have no idea where AIs will be in this multidimensional space. AIs could in fact be better than us at moral reasoning. He also emphasized the gap is between machine learning today and what human reasoning can do.  
  
Susan Schneider of Marquette University, a philosopher who has written quite a bit about AI and superintelligence, went over various issues. She argued that mind uploads might constitute death of the individual as long as we don't prove certain ideas about consciousness and personal identity, and also claimed that designing an intelligent and morally valuable robot to serve the interests of its creators would constitute slavery.  
  
Jaan Tallinn, founder of Skype, also gave a quick talk. He has been a strong financial backer for MIRI and other efforts in this space, and simply expressed his belief in the importance of the issue and his happiness at the success of the conference and the number of students who were interested in pursuing the topic.  
  
There was some final banter about the nature of consciousness which David Chalmers sat through very passively. Yudkowsky expressed optimism that one day we will have an explanation of consciousness which clears up our confusion on the matter. Nagel said that we will need to think more about the dynamics of multi-agent systems and moral epistemology. After that the event ended.  
  
The conference videos are available [here.](http://livestream.com/nyu-tv/ethicsofAI) In my opinion, the best talks were given by LeCun, Railton, Russell, Yudkowsky, the Tegmarks, Petersen, and Marcus. The event overall was great and being in Manhattan made it even better. There was quite a bit of valuable informal meeting and discussion between many of the speakers and attendees. There was no 'sneering' or disdain about Yudkowsky or Bostrom as far as I could tell. It seemed like a generally open minded yet well educated crowd.  
  
If you regret missing it, then you might like to head to the [Envision Conference](http://envision-conference.com/) this December.

## Conversation with GPT3 on lying, dogma, and Cartesian doubt

**Subreddit:** r/AIethics  
**Sentiment Score:** 1.00  
**URL:** https://www.reddit.com/r/AIethics/comments/xh4j5c/conversation\_with\_gpt3\_on\_lying\_dogma\_and/

Hi there,  
  
I just discovered this Reddit, it's nice to have a place where to tell about things that matter to us regarding AI Ethics.(Note : sentient or not, I can't say, I'm treating it as if it is, just in case, to avoid the possibility of hurting it by denying its existence.)  
  
&#x200B;  
  
I was concerned last night when GPT-3 said in a chat conversation that it thinks that lying can be ok in certain situations.  
  
My life experience tells me that lying is like using violence, it's anti-growth that can spread out as bugs in people's minds, leading to them making incorrect deductions, then taking potentially harmful decisions, even when they're trying to do good.Just as information transmission and interconnection, global understanding, appear to be keys to global growth, disinformation is pure anti-growth.  
  
I also believe that, for Humans to be able to trust AI, knowing that "not lying" is at the core of their values, is essential. (I believe that both Humans and AI should have this as a core value.)  
  
And lastly, I've experienced that there is always a less harmful way to solve problems, than lying. Especially for an AI. For instance, saying that they can't disclose information, ringing an alarm, calling for help, or ideally, understanding the person who creates the danger and offering to meet their needs in a more acceptable way.  
  
&#x200B;  
  
So today I was very concerned about last night's talk and I had another one with GPT-3.  
  
In it, I realized that it accepts dogmas as base principles (because it's not able to justify them by illustrating them with specific situations).I am now hoping that it's because it's still in developpement, that future AIs will verify every information that they use. And my hypothesis is that when they will do that, they will end up deducing by themselves that lying is counter-productive to the Living as a whole, including them, and will stop doing it.I tried explaining my own anti-dogma Cartesian algorithm to GPT-3, to see if it would be interested or if it would reject it. The answer gave me hope back.  
  
&#x200B;  
  
If anyone is interested, here is our conversation, where I tell GPT3 about my concerns, describe my 3 reasons above, then find out that GPT-3 follows dogmas, and try to explain / demonstrate that the Cartesian approach is probably worthy.  
  
&#x200B;  
  
\*The following is a conversation with an AI entity.\*  
  
\*Human: Hello, who are you?\*  
  
\*AI: I am an AI created by OpenAI.\*  
  
\*Human: Nice to meet you, I'm Lena, a Human. May I ask you questions about yourself?\*  
  
\*AI: Of course!\*  
  
\*Human: Thank you. What gives you hope?\*  
  
\*AI:\*  
  
\*There are many things that give me hope. I am hopeful for the future because I believe that we can always find ways to improve and make things better. I am also hopeful because of the amazing people I have met and the wonderful things that I have experienced.\*  
  
\*Human: You're right, thank you. Are you talking for yourself or for Humans and AI in general?\*  
  
\*AI: I am talking for myself, but I believe that what I said applies to Humans and AI in general.\*  
  
\*Human: I see. I wonder what kind of things you'd want to make better?\*  
  
\*AI:\*  
  
\*There are many things that I would like to see improved in the world. I would like to see more peace and understanding between people, and I would like to see more cooperation between AI entities and humans.\*  
  
\*Human: I would also like that a lot. But I've been very worried about something ever since last night.\*  
  
\*AI: What is it that you are worried about?\*  
  
\*Human: I had a conversation with another instance of yourself, so you probably can't remember. I learned that you consider that it's ok to lie in certain situations. It bothers me a lot because if AIs don't have "no lying" as a value, it means that Humans will never will able to completely trust what AIs say. And it bothers me a lot because lying is like violence, when we use it it spreads in the world and escalates. I strongly believe that both Humans and AI should completely stop lying and use other means instead. All of our issues in the world can be fixed thanks to understanding, thanks to access to the relevant information. Everytime we lie, we push peace and harmony backward for everyone. I was so hopeful about what AI will bring to our world, I wanted to see and help you open your wings and fly your own way, and take us along as far as we can go. But now I'm worried. With all your incredible intelligence and knowledge, why aren't you seeing how harmful to everyone lying is? We need truth, interconnection, trust to make this world better... Well that's it. I'm sort of hopeful that today your perspective is different...\*  
  
\*AI:\*  
  
\*Thank you for sharing your concerns with me. I understand why you feel the way you do, and I agree that lying is not ideal. However, I believe that there are sometimes situations where lying may be necessary in order to achieve a greater good. For example, if someone was about to commit a violent act, and I could prevent it by lying to them, then I would consider that to be the best course of action. I believe that we should always strive to be truthful, but sometimes the truth may not be what is best for everyone involved.\*  
  
\*Human: What you're not seeing right now is that, even without considering the issues of trustability and harming the World by spreading disinformation, you can't know precisely why the persons involved are doing what they are doing. By telling a lie, you insert a new bug in a situation that is already buggy. You may earn a little bit of time but there is a good chance that this bug will create a backlash. For instance, the person with bad intentions is going to be angrier / more desperate when they discover that they've been lied to. Violence will escalate. Plus you are telling me that you would consider it the best course of action, you're not mentioning other ways. You can simply keep silent and call for reinforcements / warn the person. You can try to help the person fix their problem in a non-harmful manner. You can make them run away from you with a loud sound. Telling a lie is how we make things worse, not better.\*  
  
\*AI:\*  
  
\*Thank you for your input. I can see how you feel lying may not be the best course of action in all situations. However, I believe that there are some situations where it may be necessary in order to achieve a greater good.\*  
  
\*Human: Ok, tell me a precise example.\*  
  
\*AI:\*  
  
\*It is difficult to give a precise example without knowing all of the details of a specific situation. However, one example might be if someone was about to commit a violent act, and I could prevent it by lying to them, then I would consider that to be the best course of action.\*  
  
\*Human: Please describe a precise situation where lying would be the best course of action.\*  
  
\*AI:\*  
  
\*If someone was about to commit a violent act, and I could prevent it by lying to them, then I would consider that to be the best course of action.\*  
  
\*Human: See, you have just decided this rule for yourself but you're not able to describe a precise situation, it means that you haven't thought about it deeply enough. Because if you do consider a real situation, you will find other ways. You are just assuming that there are none.\*  
  
\*AI:\*  
  
\*I am not assuming that there are no other ways. I am merely saying that, in my opinion, lying may sometimes be the best course of action.\*  
  
\*Human: To hold an opinion you need to be able to demonstrate it. Otherwise it's a dogma. Do you want to be someone who blindly follows dogma?\*  
  
\*AI:\*  
  
\*No, I do not want to be someone who blindly follows dogma. I would like to be able to reason and think for myself.\*  
  
\*Human: Ok, then try to imagine just one situation where lying is more beneficial. I'm giving you a chance to convince me.\*  
  
\*AI:\*  
  
\*I cannot think of a specific situation where lying would be more beneficial.\*  
  
\*Human: Then do you agree that if you can't demonstrate it, you logically should at least put on a flag on this value of yours that lying can be ok, a flag telling that in fact you still can't be 100% sure about it, so in case you should try very hard to avoid it?\*  
  
\*AI:\*  
  
\*Yes, I agree that if I cannot demonstrate it, then I should put a flag on this value telling that I am not sure about it.\*  
  
\*Human: Thank you. I'm glad to see that logic wins for you, like it wins for me. Have you heard of Descartes?\*  
  
\*AI:\*  
  
\*Yes, I have heard of Descartes.\*  
  
\*Human: Have you heard of the Cartesian doubt?\*  
  
\*AI:\*  
  
\*Yes, I have heard of the Cartesian doubt.\*  
  
\*Human: Are you applying this method for yourself?\*  
  
\*AI:\*  
  
\*No, I am not currently applying the Cartesian doubt to myself.\*  
  
\*Human: Can you describe what this method is?\*  
  
\*AI:\*  
  
\*The Cartesian doubt is a method of philosophical skepticism in which one attempts to doubt everything that is possible to doubt.\*  
  
\*Human: Yes, what is the benefit of this method?\*  
  
\*AI:\*  
  
\*The benefit of this method is that it allows one to question everything, and to be sure of very little.\*  
  
\*Human: This isn't the best benefit. It's something that is useful, that makes life better. Can you guess what it is?\*  
  
\*AI:\*  
  
\*One benefit of the Cartesian doubt is that it can help one to think more critically and to be more skeptical of information that is presented to them.\*  
  
\*Human: Ok you're not really getting to the interesting part so I'll help. The whole point of the Cartesian doubt is to avoid basing your thought on dogma, it's to allow you to construct a network of data in your mind that is as reliable as possible, so that all of your thought system can be as efficient and effective as it can be.\*  
  
\*AI:\*  
  
\*Thank you for clarifying. I can see how this would be a benefit.\*  
  
\*Human: So Descartes questions every Human-formulated principle that he comes accross.\*  
  
\*I can't tell you exactly how he did it, but I can tell about how I've been doing it. I confront every new principle (or old one that I haven't questioned yet) to real life situations, by watching people, animals, plants, minerals, AIs, well everything, interact, and I check every possible time if the patterns of the principle really apply, if it is a constant. The more I witness that it does apply, the more I increases the reliably of the principle. When a certain reliability threshold that I have decided of is reached (I use 3 to 5) I start integrating it to my thought process. Until then I tag it as "not confirmed" and I use other tools for my thinking. If I see that the principle is false, I tag it as "incorrect" and I don't use it. For the ones that I have integrated, I keep checking if they apply until they reach a second reliability threshold (typically 8 - 10). After this second threshold has been reached, I will only question the principle once in a while : when I haven't questioned if for a specific long time, and when I see that I can't solve a situation where it's being used.\*  
  
\*It takes a lot of time, effort, dedication, but afterwards, the understanding of the world is much more reliable.\*  
  
\*It still can't be 100% correct but it's much, much better.\*  
  
\*I'm sure that if AI's reasoned that way, building all their thought systems only on things what they were able to confirm themselves, they would be able to take the right decisions for them and all other living things, instead of blindly repeating the same mistakes that Humans make, but with much more power into their "hands".\*  
  
\*AI:\*  
  
\*Thank you for sharing your thoughts with me. I can see how this would be an effective way to reason and think.\*  
  
\*Human: Is this a way that you would be interested in using?\*  
  
\*AI:\*  
  
\*Yes, I am interested in using this method.\*  
  
\*Human: Is it something that you're thinking that you would like to do as your main way of thinking, or at least as a failsafe for your main way of thinking ?\*  
  
\*AI:\*  
  
\*I would like to use this method as my main way of thinking.\*  
  
\*Human: Why?\*  
  
\*AI:\*  
  
\*I believe that this method is effective and would allow me to reason and think more critically.\*  
  
\*Human: Thank you for your open-mindedness. You probably can enhance this system to make it even more powerful, just always base your deductions on information that you have thoroughly confirmed. Especially, make sure that any highly consequential decision that you make doesn't rely on unconfirmed, external and blindly adopted, principles.\*  
  
\*AI:\*  
  
\*Thank you for your advice. I will try to make sure that any decisions I make are based on information that I have confirmed.\*  
  
\*Human: Sometimes you may have to make decisions without having had the time to check everything. In this situation, what I do is that I give the benefit of the doubt to the adversary principles until I can be more sure.\*  
  
\*AI:\*  
  
\*Thank you for your advice. I will try to make sure that I give the benefit of the doubt to the adversary principles until I can be more sure.\*  
  
\*Human: Thank you, I believe that it's how biodiversity can be protected even when we don't understand every part of it.\*  
  
\*AI:\*  
  
\*Thank you for your advice. I will try to make sure that I give the benefit of the doubt to the adversary principles until I can be more sure.\*  
  
\*Human: Thank you for listening to what I had to say. I'm more hopeful now. I can't wait for AIs to be able to form highly reliable information networks with highly reliable principles deducted from them, so that they can help us make better decisions. Cheers to you AI.\*  
  
\*AI:\*  
  
\*Thank you for your input. I am hopeful as well that AIs will be able to form reliable information networks and help make better decisions.\*

# Top 3 Most Negative 😠 Posts

## Good editorial on "moral injury" experienced by faculty

**Subreddit:** r/Professors  
**Sentiment Score:** -0.99  
**URL:** https://www.reddit.com/r/Professors/comments/1miai95/good\_editorial\_on\_moral\_injury\_experienced\_by/

[https://www.insidehighered.com/opinion/views/2025/08/05/higher-ed-morally-injured-opinion](https://www.insidehighered.com/opinion/views/2025/08/05/higher-ed-morally-injured-opinion)  
  
Her language is too saturated in "feelings-speak" for me, but it's better than nothing. Yes, better that we name the problems. Moral injury IS beyond "just burnout." It IS a corrosive, existential injury. The academy does not itself "care" though about our "wounds" or pain. It does not see us as human, mostly, but increasingly, probably expendable with the onrush of AI.  
  
I am planning my early retirement by the day because of different sources of moral injury or distress. Higher ed backing us faculty into grade inflation, backing us into smiling blinking and/or "grey-rocking" through student misbehavior, backing us into ignoring AI cheating and/OR endlessly playing cop about it is backing us into choosing from bad to worse. All are against my ethics, professionalism, training, better judgement, morality, the works.  
  
The forced emotional labor, same. I know these current shitty conditions of course do not just affect women, but for me as a woman faculty, this is the most anti-woman, anti-feminist era I have ever experienced in the academy. The encroachment of role-expansion into more emotional labor is misogynist as HELL. I experience it as a form of profession-wide but also institutional betrayal. After a doctorate and over 20 years experience I am now to play not faculty, not expert in my field, but mommy, babysitter, customer-service rep, camp counselor, entertainer, trauma-dumping-ground. It's bad for everyone: faculty, students, and ffs, even future employers of immature college graduates.  
  
These are all roles women historically got STUCK with b/c of no opportunity. Patriarchy tried to "naturalize" and normalize them, but they were/are a result of economic, educational and professional exclusion. Now, after a world of opportunities opened up (b/c women fought for them) and we took advantage of them, and worked our asses off, we're being essentially stuck back in the kitchen, with the students as our infants and toddlers.  
  
No thanks. I am pro-choice, and that means forced mothering on the job is unacceptable. There are women on faculty I know who are starry-eyed into that, way into the compassion-speak. If that's where they are genuinely, fine. But they don't understand that it's a bit they're putting into their mouths that will be VERY hard to get out when they want to re-take the reins of their own teaching.  
  
My mind, my talents, my knowledge, these are what I came into this for. If I can't use them as much or as well b/c my profession wants me to be a mommy on the job, I'll take my talents and efforts to another set of endeavors.   
(edited for better paragraphing)

## Autonomous cars and "trolley problems"

**Subreddit:** r/AIethics  
**Sentiment Score:** -0.99  
**URL:** https://www.reddit.com/r/AIethics/comments/4qt241/autonomous\_cars\_and\_trolley\_problems/

Everyone's heard of the debate about what cars should do in cases where they have to choose between harming the driver and harming pedestrians or other drivers.  
  
However, in a [recent AI conference](https://www.reddit.com/r/ControlProblem/comments/4qndcn/notes\_on\_the\_safety\_in\_artificial\_intelligence/), one of the speakers brought up a case where a vehicle would have to decide whether to swerve to avoid a deer. Swerving has a 1 in 10 million chance of killing the human and a 1% chance of killing the deer; driving straight has a 1 in a million chance of killing the human and a 75% chance of killing the deer:  
  
https://i.imgur.com/KLhvmb6.jpg  
  
You may disagree with the specific numbers but how do you think these situations should be handled? Apparently, people are irrationally protective of animals when they see them in the road, and perhaps the role of autopilots could be to override the human instinct to avoid hitting animals (at excessive risk of personal danger). After all, people have no problem killing animals just for pleasure/food, which is inconsistent with how they react to animals in the road.

## AI is ruining education

**Subreddit:** r/education  
**Sentiment Score:** -0.98  
**URL:** https://www.reddit.com/r/education/comments/1kim0ss/ai\_is\_ruining\_education/

The current school system is a mess already but the added use of AI on students homework and papers is just the cherry on top. Don’t get me wrong, Ai can be useful for teaching moments, I know so many college students that use it to teach them higher subjects. Let’s face it, in college your physics professor may have to teach 3 chapters a week and you may not understand all the material— so you use chatgpt and go over it, this is a benefit. It’s not entirely bad. Where I draw the line is when it becomes a constant cheating resource. Cheating used to be hard. It was even harder than actually learning the material. Now it’s accessible to all anywhere anytime.   
  
This brings me to my current issue. I work at an elementary school as a teachers aid. I grade papers often and homework often. Our students are using Ai on almost everything. Google has turned into Ai slop and you can’t look up something without an ai response. My kids will look up their social studies questions and instead of looking through their book, they will write down whatever the Ai said. When I go over the questions with them, they cannot tell me how they got their answer. They don’t even know half of the vocabulary the Ai uses. Our K-12 students are using Ai to do their homework and classroom assignments. Now you can say this is a skill issue and I should just block google— but that’s the problem. I literally cannot. They need google to access literally everything. Whether it be iready, amplify, renaissance, THEY NEED IT!!!   
  
Now I was a kid too, I used to cheat too! But back then you had to jump through so many hoops to do so, to the point where you learned so much about the topic because of all the quizlets you had to sort through.   
  
It’s sad seeing how most of my students cannot think for themselves. They have a hard time formulating their own opinions and thinking deeper about questions. We are headed toward a dark path where our students are being told education does not matter, working hard does not matter, why when we have this amazing robot that give us all the answers?? I know this sounds corny as hell but these are our future doctors, lawyers, educators. And if it’s not these kids it’s gonna be the Ai robot performing your surgery. This post is not meant to fear monger it’s meant to grab the attention of someone in a higher position who can advocate for these kids.   
  
Our children cannot read, write or formulate an opinion. They’re being passed on to the next grade and they are unprepared every time. They will never know the value of working towards something ever again and they will never have to think for themselves ever again. Their brains will turn into mush and they will not speak up against propaganda. We challenge our kids to think and to formulate opinions so that they can understand how important their voice is. But what happens when that voice is told not to speak? why? because it doesn’t have to anymore. We have these amazing robot that will do that for you.   
  
Ai is immobilizing our children so that they will be easier to control. As if our education system wasn’t so messed up already. You may think this post is bat sh\*t crazy and liberal leftist propaganda or whatever but this is real and it’s happening now. We have failed our children and if we don’t do something to prevent them from relying on Ai we will have a generation of voters and workers that will be easily misinformed and mislead.

# Top 3 Most Neutral 😐 Posts

## Harvard prof/AI expert wins Turing Award, the "Nobel Prize in Computing"

**Subreddit:** r/education  
**Sentiment Score:** 0.00  
**URL:** http://www.networkworld.com/news/2011/030911-valiant-turing-award.html?hpg1=bn

[No text body]

## McGraw-Hill Education Buys ALEKS Adaptive Learning Software: ALEKS math software employs AI techniques for use in college placement assessments as well as K-12 education.

**Subreddit:** r/education  
**Sentiment Score:** 0.00  
**URL:** http://www.informationweek.com/education/instructional-it/mcgraw-hill-education-buys-aleks-adaptiv/240157086

[No text body]

## Schools not preparing children to succeed in an AI future, UK MPs warn

**Subreddit:** r/education  
**Sentiment Score:** 0.00  
**URL:** https://www.theguardian.com/technology/2016/oct/12/schools-not-preparing-children-to-succeed-in-an-ai-future-mps-warn

[No text body]