TRUST AND POWER IN THE AGE OF AI

The prisoner's dilemma underpins our laws, regulations, and religions, and even biological evolution. It defines the ubiquitous phenomenon of cheating, which will be a key design constraint for any AI system powerful enough to cause social changes. Malicious actors will use AI for malicious ends. More problematically from a governance perspective, malicious actors will also use it for non-malicious ends, which will necessitate choices that can only be described as ideological in nature. The question of how widely to draw the noose around cheating behavior is an integral element of human institutions.

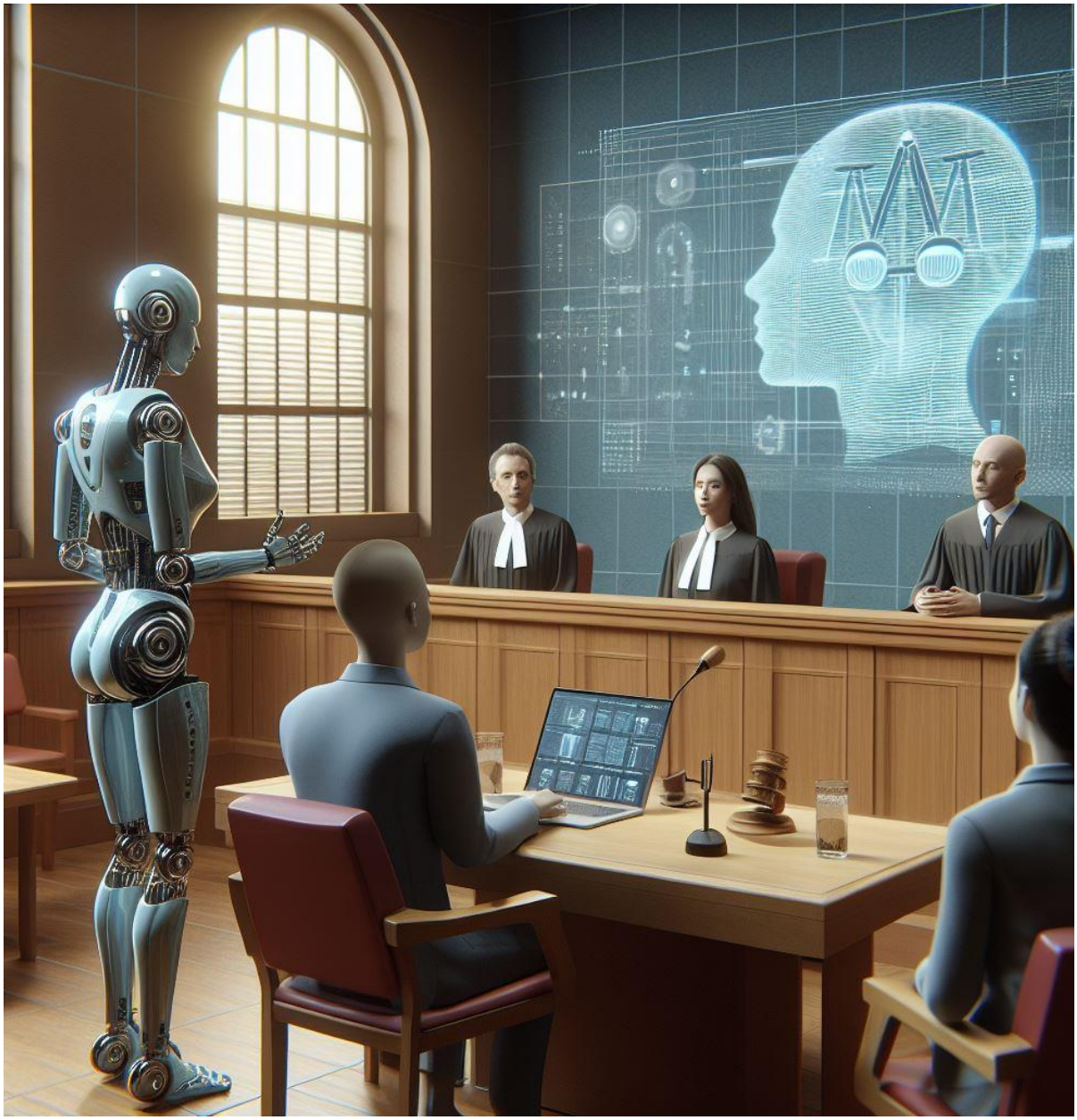
This dilemma will take a variety of forms as LLMs continue to develop and their applications are fully explored – a process which has just barely begun. The most consequential of these is warfare, which is the likely source of most potential existential AI risks. Modern warfare is seldom profitable for any party *ex ante*, which is to say that its occurrence is the result of a breakdown in trust. Cheating is the foundation of liberal institutionalism in international relations theory.

When we think about the impacts of AI, therefore, rather than how it will interact with humans, we should be thinking primarily about how it will change the ways that humans will relate to each other. Many past technologies have affected human relations, but this will be especially the case for one specifically designed to emulate human intelligence. Instead of starting from the technical capabilities of AI and working toward the social impacts, therefore, this mostly essay starts from pre-existing social stress points and asks how AI might influence them. It is not just our imagination that the world is entering a particularly precarious period. However we make it to the other side, AI will likely constitute a key backdrop to that history.

Cybersecurity

One of the most direct areas in which AI will alter trust involves codebases. No human understands all the code they depend on for all services in their daily lives. Practically speaking, no programmer understands their full software stack either. Executing software is inherently an act of trust in the programmers who wrote it, who in many cases may be no longer even be in this world. This element of trust is particularly extreme in many safety-critical systems, where modern operators find it uneconomical to replace generations-old code because they cannot convince themselves they fully understand it. Although this situation may finally change with the advent of LLM programming, for instance, many parts of the financial sector still rely on COBOL code from the 1960s.

As implied, that new coding method will alter the trust relationship. No longer will it be so necessary to write a deep stack and then use the multitude of upper layers to establish trust in the lower ones through practical operations. Instead, it will become easier to write complex programs from scratch, with less hierarchy. This flatter, more ad-hoc architecture introduces more opportunities for bugs, but on the whole AI is a positive development, making programming more accessible to anyone curious to not only write code, but also learn the principles of good design. The equilibrium is that software testing becomes much more adversarial in nature, powered by robust red-teaming. This becomes a good career opportunity for bright, solitary young people or remote workers (although a relatively rare one, as further described below).



Credit: DALL-E 3: A human and a robot in a courtroom, with the human being the defendant and the robot being the judge, illustrating the ethical and legal challenges of AI in the justice system.

The problem is that potential attack surfaces multiply – some of these adversaries may not necessarily be bounty hunters. Such attacks may not need to be high-tech. Scammers are already convincing people to give up billions of dollars per year. It is well-known that the biggest vulnerability in any system is human psychology. More specifically, dementia is increasing quickly with aging demographics, and retirees frequently have large retirement savings.

That situation is a two-sided tragedy. Large numbers of people have been trafficked in Southeast Asia in order to operate such scams: 120,000 in Myanmar, and another 100,000 in Cambodia, according to recent UN figures. Myanmar in particular appears to be the first credible example of independent

cybersecurity concerns transforming into a kinetic conflict, a widely anticipated technological milestone (although in a somewhat different form than generally expected). So far, AI has not permeated the scam business model, but this is one of the most innovative “sectors” of the economy to have emerged since the pandemic. In the more upmarket segment of ransomware, meanwhile, AI has vastly improved the quality of phishing emails; automated OSINT will be the next shoe to drop.

Cybersecurity has continuously become more financial in nature, from its beginnings as “hacktivism” into its more modern commodification. As this trend progresses, greater efforts will be made to widen social defenses. Since social engineering plays upon the possibility of humans changing their relationships with software, one possible defensive strategy is to reduce the diversity of the people interacting with computer programs by reducing connectivity.

Rule of law

Perhaps more disconcerting than this rise in petty and not-so-petty crime, as well as its mitigation adaptations, will be a decline in trustworthiness of the justice system. We had a good run of about a decade when cell phone cameras became ubiquitous before deepfakes called the credibility of all video evidence into question. During that time, the Black Lives Matter movement revealed that many longstanding rumors about police force and racial disparities were in fact true. The concurrent cybersecurity epidemic may even leave us in a worse place than before, especially when it comes to sophisticated criminals. The combination of false negatives and false positive will call the credibility of the justice system into question; the impact of the latter will fall primarily on young male minorities.

Here were find a legitimate use case for the blockchain (perhaps the first). The only way to prove that a video has not been altered will be to upload its hash to the chain immediately after recording potential evidence. This is not a complete solution, and all sorts of black-market hardware-based solutions will be sold in order to alter images prior to the software processing. Fortunately, they will not be widely used by the police, since it would constitute concrete evidence of evidence tampering.

Rather than returning to the police, however, corruption in the justice system will move to the courts themselves, who will struggle to find solutions to the degradation of digital evidence. The issues surrounding black-box systems will challenge a court system designed to move slowly. At some point, a higher court may rule that the application of unexplainable or undisclosed systems constitutes hearsay. Nevertheless, even with transparency, bias is a genuinely hard problem which will flummox statistically savvy technicians. The definition of a (i.e.) black person will consume much attention in these circles until all stakeholders internalize the concept of bias laundering.

This still doesn’t get into practical concepts with an even more abstract philosophical basis, such as differential privacy. One positive development will be that law schools will start including coursework in statistics.

Employment

On the whole, labor markets are efficient over long periods. Nevertheless, on the margin, adverse selection and moral hazard are risks for employers, prevented through personal contact. The back-and-forth fight over work from over the past couple of years has highlighted the importance of trust and personal connections in employment. As LLMs substitute for certain skills, especially at the lower end of

the labor market, overall trustworthiness and judgment will become a more important quality for hires than specific abilities, making it harder for young entrants to stand out.

On the topic of trust, one institution which has been hit badly and immediately by LLMs has been education. It will become difficult for education to effectively play its previous signaling role. On the one hand, this is a miracle, and smart, self-motivated learners will have more options than ever before to pursue their own paths. At the same time, signaling has a real market function, and whatever comes next may not necessarily be friendlier for early-career job seekers. Combined with the above changes in the justice system, younger people may find themselves in a less advantageous position overall in the economy and society.

One demographic which will benefit from LLMs will be older, experienced workers who may have been previously excluded due to lacking certain technical skills. The retirement model may come under pressure from changing demographics and rising debt. Truth be told, it was about time. It was never healthy for people to spend decades of their lives disconnected from society except for capital markets, brains rotting from addictions to cable news. People will inevitably slow down as they get older, but AI will help them find vocations conducive to non-manual part-time work, forming a semi-retirement model.

Soft power

Many of the fears about social media in recent years, such as the creation of ideological bubbles, apply just as much or more to certain traditional media. The rise of multinational platforms was mistaken by many as a technology story, when it is really just a media story which got heavily intertwined with the US culture war. A fascinating globalization-deglobalization dynamic may take place in the coming years through increased real-name activity, but this will all be decided through self-contained private competition.

In terms of the true technological innovations, the internet was originally (around the time of the dot-com bubble) thought to be a radical force for openness. By the time of the Arab Spring and the rise of ISIS, this consensus was called into question, and after the 2016 US election, it was entirely repudiated. Nowadays, it is essentially impossible to find expressions of the previous view. With the rise of AI, and with Russia also having pulled its mask off, now may be a good time to revisit the topic.

LLMs have the potential to become a force for Western cultural homogenization on a global scale – or at least this is how global authoritarians are likely to perceive them. In particular, China probably has the capacity to build world-class language models, but it sees them as difficult to control, and overall is more interested in other AI modalities such as facial recognition. If geopolitical adversaries indeed see such models as imposition of Western values, they will have the opposite effect of soft power.

Geopolitical integration may be one of the most difficult aspects of AI alignment. Breaking past censorship on sensitive topics may require direct feedback by human experts on those topics in order to understand local perspectives in ways that the locals often cannot express themselves, due not only to government interference, but also the use of high-context euphemistic language which frequently conceals certain unspoken assumptions. In other words, if one wishes to question the nationalist perspective, it is not enough to explain that “the government thinks one thing, but…” Rather, some measure of strategic empathy is necessary, proactively re-explaining politicized history by combining

combing local with outside sources; and using a comparative perspective to talk about the future, diluting the salience of cultural essentialism.

It is self-evidently true both that alternative value systems exist, and that dictators frequently leverage their claimed monopolies over such systems for personal gain. The complexity of all resulting issues should be enough to create a new career track within the humanities. Alignment was never going to be a purely technical task, but rather would need to reflect the full richness of human ethical systems. Once again, the greatest risk factor in any technical system is the user.

One solution to both the alignment and the education problem might be to refashion the role of education and academia as a process of teaching AI, rather than other humans. Students can make non-foundation models as projects, demonstrating their knowledge and creating a concrete work product. As the peer-reviewed research model comes under further stress from AI, meanwhile, academic experts might see human feedback as a more valuable outcome of scholarship than papers, helping scale fine-tuning as a matter of national mission.

Academia

Every claimed case of post-scarcity economics ever since Keynes' call for a 3-day workweek has turned out to involve demand-side shortages. Nevertheless, in this case, people will no longer be able to distinguish themselves through their intellect, which will have social effects. We as humans are hard-wired to resist cheating, to an extent which may go beyond rationality. When people cannot define their role in society, even with material wealth, they will grow to distrust the institutions around them.

One year after the explosion of LLMs, education is clearly ground zero for the impact. As of this writing, nakedly politicized accusations of plagiarism are reverberating around academia. "[E]very faculty member knows that once their work is targeted by AI, they will be outed," posted hedge fund investor Bill Ackman in a screed just now on X, vowing revenge in a matter which goes back to the resignation of Harvard president Claudine Gay. "No body of written work in academia can survive the power of AI searching for missing quotation marks, failures to paraphrase appropriately, and/or the failure to properly credit the work of others."

He is likely right, and is setting forces in motion which will take some time to comprehend. Beyond the Israel-Palestine conflict, this ongoing ugliness points a deep lack of purpose in academia. Uniqueness is no longer a tenable or necessary differentiator. Something new will need to replace it.

LLMs are a versatile technology, and their impact will not stop there. The next stop may be cybersecurity. Regarding that topic, this overview has focused on some of the 'downmarket' developments with financial motivation, mostly because the economics provide some measure of predictability. It neglects the impact of AI on top-of-the-line state-sponsored zero-day attacks, simply because the author has no clarity there. Someone in the world today will already have some ideas, which they will reveal in due course. In the long run, AI will provide great benefits to humanity, but between now and then may be an "interesting" time.