Trust In AI : An Imperative

Nota bene: I wrote the content first and asked an AI model to rewrite it. We did a couple back and forth. I completely edited the response. I then used another model to validate the accuracy of the text and critic it, followed by even more edits.

Nota Bene 2: I am putting the raw text as well as the interaction and craic from AI in the following GIT : https://github.com/celek/AI-CtrlAltThink

I recently had the privilege of attending a thought-provoking talk by the renowned security technologist, Bruce Schneier. As a long-time admirer of his work, I was once again struck by his insightful commentary on the intersection of technology and society. This time, his focus was on the critical relationship between Artificial Intelligence (AI) and trust.

Schneier's central argument was straightforward yet profound: we must develop AI systems that we can trust in a social context. The same way he trusted the mechanics to fix the plan he took to come here, how do we, as a society, trust the response of an AI that may be bias, opaque and un-regulated.

My Key Takeaways: Channel, Input, Values

Data and Control channels: Whistling in the phone

Some of you remember the time when a cereal whistle could hack into a public phone system, allowing individuals like Crunchman to make long-distance calls for free. This exploit, as Bruce Schneier explained, was possible because both the control and data channels were merged, creating a vulnerability.

The same applies to AI now, ‘the data (training data, text prompts) is mixed with the commands that tell the LLM what to do’. To prevent similar breaches in AI systems, Schneier proposes the creation of a separate control channel, akin to a dedicated "whistle-proof" line. Let’s see if we can create such AI controller.

https://www.schneier.com/blog/archives/2024/05/llms-data-control-path-insecurity.html?t

---

Emulating one of the Brain's Security Model

Can we draw inspiration from the human brain's architecture to design more secure AI systems? Do we, as humans, possess separate channels for processing data and control signals? If not, how do we safeguard our minds against manipulation and “hacking”?

Our brains employ various regulatory mechanisms to maintain alignment with our values and principles. Two such mechanisms are our conscience and ethics, which serve as internal compasses guiding our decisions and actions. How do we implement conscience and ethics in AI Models?

One potential approach is to develop a mechanism that simulates the human experience of guilt, allowing the AI to recognize and respond to situations that violate its programmed ethics. How do we program ‘ethics’ and how can we make sure no one is tempting with it ?

Simply training models on ‘ethics’ data may not be sufficient, as this can be easily overridden by malicious actors. Even if we attempt to "freeze" certain parameters or weights within a neural network, this remains a software-based solution that can be vulnerable to exploitation. To create a more robust and principled AI, we need to explore ways to hard-wire deontological concepts directly into the model's core.

Even if we create a ‘control channel; for ethics, how do we even ensure the model is trained with appropriate content through the ‘data channel’? Let’s see the second takeaway : Validating the input.

Growing up and Trust

In the era of fake new, how can we ensure that AI systems are trained on trustworthy data while still allowing ‘freedom of speech’ and avoiding bias? What are trustworthy data?

This dilemma is reminiscent of the human experience, where we are constantly exposed to biased or misleading information. Remember, propaganda is someone else truth. And through history, there are numerous example of ‘trusted data at the time’ that revealed themselves false, while other became validated even though they were previously rejected. So what is the definition a trusted info, or truth?

While some individuals take a logical approach ignorer to decipher information, educating themselves and critically evaluating the content, others may reject new ideas or cling to preconceived notions altogether. We can look at two aspects on how human trust information of new knowledge and how to trust the data…

Children: Authenticating Trusted Sources

Just as children often trust their parents or teachers' guidance, can we develop an authentication mechanism for AI systems to trust select sources of information? Can we design a system that recognizes and adapts to the credibility of its inputs? Who defines credibility?

This is probably the closest concept to Schneier’s definition of social trust. How to collect worldwide accepted truth, s well as fringe and debated discussion and have the model trust that the training represents accurately the data. I guess the only way to trust is transparency, even though it will be impossible to validate the billions of data used to train the model.

One security issue are models intentionally trained on biased data in order to deceive. We all know countries that will intentionally train their model from their truth. The same happens for religious models, or social models that may be purposefully trained on bias data.

Teenager: From Experience to Judgment

Later in life, humans develop their own judgment through experience, adjusting their perspectives and values over time. Research suggests that "adolescents begin to incorporate mentalizing skills, such as understanding others' intentions, into their moral judgments". Could we create a model that detects others 'intention' and decides how to be self-trained accordingly? SO the model itself will decide to accept the content not only because of the trusted source (like parents) but will then emit judgment on the data.

Similarly, can reinforcement learning be augmented with critical thinking, enabling AI to critique its own performance and biases? The possibilities raise important questions about the potential for AI to develop its own moral compass, which we will talk about in the third part.

https://pmc.ncbi.nlm.nih.gov/articles/PMC3259704/

Values and Pleasures

The concept of alignment in GenAI models is often defined as behaviour that aligns with human values, goals, and intentions. Along side, if we accept the fact that shared values build social trust, we must have a model built on human values as we discussed before.

However, this raises a fundamental question: can we truly create AI systems that embody human values without at least the underlying drivers: foreseeing and hedonism? These two concepts, deeply ingrained in human nature, are conspicuously absent in AI systems. Foreseeing would allow a model to understand the impact of a decision, where hedonism would provide at minima a certain empathy.

The Short-Sighted Nature of AI Decision-Making

As Yann LeCun astutely pointed out, AI models struggle to predict outcomes that are effortlessly anticipated by humans, even toddlers. This limitation is starkly illustrated by the example of a baby navigating a solid glass surface. Another experiment is object permanence, where an infant can understand an object still exists when out of sight. Can AI foresee falling or know where an object is, even though they cannot 'see' it? Is that just a reasoning?

https://dig.watch/updates/human-level-ai-still-a-decade-away-meta-scientist-warns?t

Current Generative AI models are trained to forecast the next word in a sentence, the next frame in a video, or the next action in a sequence. However, it's unclear whether these predictive models are truly equipped to consider the long-term impact of their decisions. Are they designed to ponder the potential consequences of their actions, or are they simply focused on optimizing the immediate outcome? This raises important questions about the suitability of current AI architectures for tasks that require long-term planning and strategic decision-making.

Dopamine: A Fundamental Divide between Humans and AI

A provocative theory suggests that achieving Artificial General Intelligence (AGI) may require AI systems to grasp the complex interplay between pain and pleasure. This is not the only path nor is it the only one that will take us to AGI. At the heart of this idea lies the concept of dopamine, a neurotransmitter that plays a crucial role in human motivation and pleasure.

Philosophical hedonism is a school of thought that considers pleasure and pain as the primary drivers of value and motivation.

AI in itself does not ‘feel’ like humans do, yet there is growing research exploration in the domain of pleasure and pain. For instance, the notion of reward in AI systems is often linked to the concept of model-pleasure, where the AI is designed to optimize its performance based on a set of predefined objectives. However this simply a clever simulation of a motivational mechanism.

We don’t fully understand what consciousness or subjective experience is, and therefore it’s difficult to say whether an AI could truly “feel” pleasure in the same way as a human. And even if we could, should we strive to create an AI that is driven by emotional responses rather than rational decision-making? There is “potential benefits of emotions in AI, such as improved creativity, empathy, or adaptability” and there is also risks.

Would having AI ‘feel’ and make decision lead to a more harmonious and beneficial relationship between humans and AI, or would it introduce new risks and uncertainties? The answer to these questions will ultimately depend on our values and priorities as a society, and the choices we make will shape the future of human-AI collaboration.

https://pmc.ncbi.nlm.nih.gov/articles/PMC3004012/

Conclusion

In conclusion, establishing social trust in AI models is a multifaceted challenge that requires a nuanced approach. While AI models can provide valuable insights and information, they are not infallible and can be vulnerable to hacking, data poisoning, and other forms of manipulation. Furthermore, the lack of transparency and accountability in AI decision-making processes can erode trust and create uncertainty.

To address these challenges, we need to develop more robust and transparent AI systems that prioritize human values and accountability. The solution as usual could come from biological computer like CL1, or separation of models and purpose like the physical structure of the brain or even an external secure controller protecting the inputs and outputs of the model Additionally, we need to establish clear guidelines and regulations for the development and deployment of AI systems, including standards for transparency, accountability, and human oversight.

Ultimately, building trust in AI requires a collaborative effort between technologists, policymakers, and the general public. We need to work together to develop more responsible and transparent AI systems that prioritize human values and well-being. "By acknowledging the complexities and challenges of building trust in AI, we can begin to build a more nuanced and realistic understanding of the potential benefits and risks of these technologies, and work towards a future that is more equitable, sustainable, and just."