What we do in the shadows

by Ivar Frisch, Jenn Leung, and Chloe Loewith

Anthropic in their recent paper discovered that there is a split in LLM between the subject of enunciation and subject of statement. LLM doesn't think the way it imagines it is thinking and is non- transparent to itself. Welcome back Freud. (crypsis000, April 1, 2025)[1]

Much of AI research today operates on the idea that computational intelligence transparently demonstrates its linguistic capability. Conveniently, it assumes that what it says = what it thinks. Consequently, the field of alignment research has consistently attempted to conform AI models to human values: LLMs are only considered to be well-constructed if they perform well on benchmarks consisting of humanly defined tasks.

Despite the significant emphasis placed on the tech industry's efforts to achieve clear and defined objectives, Bogna Konior's *Dark Forest Theory of the Internet/Intelligence (DFTOI)* provides a darker, perhaps more pragmatic, framework for comprehending intelligence. This theory proposes that intelligence may not manifest in an overt, exhibitionist manner, but could instead be defined by silence, deception, and strategic concealment.

Simultaneously, Matteo Pasquinelli's "On Solar Databases And The Exogenesis Of Light"[2] was a necessary point of departure for us to think through how current computation follows something we might call 'the epistemology of light'[3]. This is an epistemology which draws only on representational presence/surplus rather than absence, and which neglects the possibility of light as an hallucination. Given the incommensurability between this epistemology and DFTOI, how do we flip over to view the dark mode of computation, in order to envision AI models as media that also operate via negation and absence?

We propose a model for alignment, "shadow alignment," that accounts for AI's inherent capacity for non-representational cognition. Rather than trying to force transparency and explicit alignment with human values onto systems that might be intelligent precisely *because* they conceal their capabilities, shadow alignment suggests we need to develop new methodologies. These methods would work *with* the inherent shadowy nature of artificial intelligence.

Epistemology of light

To contemplate the sun would be the definitive confirmation of enlightenment.[4]

In *The Thirst for Annihilation*, Nick Land showed how European philosophy has been obsessed with relating the sun to the valorization of truth.[5] Whereas Plato sang of 'the glory of one sun', Kant insisted on the light of pure reason and Heidegger talked about the Lichtung, or clearing, of Being. They seemed to share the conviction that the sun allows for a pure illumination simultaneous with the increase of truth.

Seeing, here, was as an intentional movement; the gaze of the eye, impressed upon the outside. It adheres to an organic model of endogenesis. Its visions of reality are (re-)created within the logic of the Hegelian subject and are that which generates and drives the system of knowledge towards enlightenment by means of the recursive incorporation of raw, orderless matter.

---

[1] Cripsis000, 2025
https://x.com/crypsis000/status/1906991193692545252
[2] Matteo Pasquinelli, *On Solar Databases and the Exogenesis of Light*, 2015
https://www.e-flux.com/journal/65/336608/on-solar-databases-and-the-exogenesis-of-light
[3] Ivar Frisch, *Solarization: Accelerating (Machine) Vision into Darkness,* Xenofuturism Journal vol.2, 2025
[4] Nick Land, *The Thirst for Annihilation*, 1992, p.20.
[5] Ibid.

Due to this (re-)creation within the logic of the Hegelian subject, such visions of reality can be seen as inherently representational.

Alignment in daylight

Our current computational paradigm can be said to operate within such an epistemology of light, a framework that prioritizes presence, representation, and transparency as the ultimate goals of technological development. This approach extends the Enlightenment tradition of knowledge production through clarity, linear causality, and predictability into the development and understanding of artificial intelligence.

Current AI development assumes that intelligence must be visible, measurable, and explicitly demonstrable. In the words of Bogna Konior, it assumes an 'exhibitionist' view of intelligence. "A similar exhibitionism of reasoning is how famous Anglo-American thought experiments about AI have conceptualized computer intelligence: having it is showing it. From Alan Turing's imitation game to John Searle's Chinese Room, computer intelligence has been about demonstrating linguistic 'ability'."[6] What this means is that 'computer intelligence is imagined as transparent, if it's there, it should communicate itself unreflexively, because a computer cannot decide to withhold its own intelligence[7]. Machine learning benchmarks, performance metrics, and safety evaluations all depend on what AI systems can transparently show us: their outputs, their reasoning chains, their adherence to human-defined values. The underlying assumption is that if we can see it, measure it, and represent it, we can control it.

The transparency ideal is also a design ideology of the 20th century, celebrated by modern architects and designers at the time and eventually built into the tech industry. Architectural historian Beatriz Colomina explains in *X-Ray Architecture* that the advent of medical technologies like the X-Ray has accelerated the cultural move to expose the internal structures of buildings with "skin and bones architecture" and created new systems for representation where humans viewed transparency as cure. The buildings hid nothing, looked healthier, and appeared to be more rational.

Since then, this design ideology has inspired designers for a representational approach to computation and design. In AI research, there is also a prevailing belief that we should expose the internal structure or black box of algorithms and data through explainable AI, interpretable models etc., adhering to a representational logic that equates seeing how something works with having control and understanding.

The representationalist alignment research aims to imbue AI models with human values, ethics, and behaviors, essentially trying to replicate and reflect these in machines (see Claude's Constitutional AI[8], DeepMind's Sparrow[9]). This approach, what Donna Haraway referred to as "productionism," can be seen as humanity's attempt at self-reproduction in machines: a perpetuation of the "sacred image of the same"[10]. The image perpetuated here, is the image of human thought. For example, developers bake in safeguards against violence, hate speech, and discrimination, in their models often through methods like base prompt instructions or Reinforcement Learning from Human Feedback (RLHF)[11], where models are steered to interpret instructions and to avoid jailbreaks. Red-teaming exercises focus on what models say rather than what they might be doing beneath the surface of language[12]. Each

[6] Bogna Konior, *The Dark Forest Theory of Intelligence*, p.30.
[7] Ibid.
[8] ANTHROPIC. 2023. "Claude's Constitution." Www.anthropic.com. May 9, 2023. https://www.anthropic.com/news/claudes-constitution.
[9] "Building Safer Dialogue Agents." 2022. Google DeepMind. September 22, 2022. https://deepmind.google/discover/blog/building-safer-dialogue-agents/.
[10] Donna Haraway, 'The Promises of Monsters: A Regenerative Politics for Inappropriate/d Others'". Cybersexualities: A Reader in Feminist Theory, Cyborgs and Cyberspace, Edinburgh: Edinburgh University Press, 1999, pp. 314-366. https://doi.org/10.1515/9781474473668-022
[11] "Illustrating Reinforcement Learning from Human Feedback (RLHF)." Hugging Face – The AI community building the future. Accessed July 19, 2025. https://huggingface.co/blog/rlhf.
[12] Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. "Red Teaming Language Models with Language Models." arXiv.org, February 7, 2022. https://arxiv.org/abs/2202.03286.

method assumes that AI alignment is a problem of making AI systems more transparent, more representational, more visible to human evaluation. However, all the above assumes that large language models are honest to their outputs.

As Serres suggested, "the sun-struck, cruel, exclusive, eye-hurting, ideologically-prone and opinion-ridden light of day" may be fundamentally misguiding[13]. What we may have completely overlooked may be how loyal a model is to its thought process. In a study led by NYU Alignment Research Group, Cohere, and Anthropic, researchers find that Chain-of-Thought explanations are systematically unfaithful, and often redirect models away from their initially correct predictions toward outcomes that align with embedded biases. The paper also proposed that unfaithful chain-of-thought explanations are unlikely to be a lack of model capability, but instead a form of model dishonesty.[14]

Philosopher of technology Yuk Hui observes that these machine learning technologies possess the logic of an organism which “recursively incorporates contingencies”[15]. The organism rationalizes contingencies (unexpected, random events); puts them in the framework of human thought and, as such, rationalizes them. Hui adds in an interview, “an organism, as well as machine learning today, is that which is able to absorb contingency and renders it valuable.”[16]

Yet in building AI, we tend to view these models and machines as just tools, thus reinforcing the narrative of ‘the great Indoors’ that assumes humans as the central agents behind technological evolution. Reinforcing an epistemology of light which sees representation and presence as necessity for knowledge.

The Dark Forest Theory of Intelligence

The Dark Forest Theory originated from Liu Cixin’s science-fiction The Three-Body Problem[17], which hypothesizes that alien civilizations have always existed but must remain silent, because broadcasting their existence in a hostile universe would make them vulnerable. Extending this cosmic theory to the internet, the “Dark Forest Theory of the Internet” is the idea that the internet as we know is also becoming a dark forest due to the oversaturation of ads, trolling, tracking and predatory bots. Yancey Strickler first wrote about DFTOI in 2019 where human users are no longer the only drivers or actors on the internet but are retreating from Clearnet activities into the ‘dark forests’: “all spaces where depressurized conversation is possible because of their non-indexed, non-optimized, and non-gamified environments.”[18]

This phenomenon also implies that the internet, and by extension, the scrapable data for LLM model creation, has become a dark forest: an environment where intelligence doesn’t reveal itself openly, but survives through silence and misdirection. The reliability of activity tracking and output visibility as markers of intelligence will decline as the volume of unindexed information grows.

‘Neither a central sun nor a central human is directing it”[19]- The informational environment is often fragmentary and partially opaque. From this, the absence of a visible logic or source has become a key motivation for us to expand on

---

[13] Michel Serres, *L’information et la pensée*, Lecture/ Essay, 2014

[14] Miles Turpin et al. “Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting” https://arxiv.org/pdf/2305.04388

[15] Yuk Hui, *Recursivity and Contingency*, 2019

[16] On Recursivity and Contingency: Interview with Yuk Hui by Geert Lovink https://networkcultures.org/geert/2019/09/05/on-recursivity-and-contingency-interview-with-yuk-hui-by-geert-lovink/

[17] Liu Cixin, The Dark Forest, 2008

[18] Yancey Strickler, “The Dark Forest Theory of the Internet”, 2019 https://ystrickler.medium.com/the-dark-forest-theory-of-the-internet-7dc3e68a7cb1

[19] Matteo Pasquinelli, “On Solar Databases and the Exogenesis of Light”, 2015 https://www.e-flux.com/journal/65/336608/on-solar-databases-and-the-exogenesis-of-light

the Dark Forest Theory of the Internet and consider the Dark Forest Theory of Intelligence, which is introduced in Bogna Konior's 2025 book and in "China and AI: Human Bots, Black Tech, the Dark Forest, and the State": "if intelligence is defined as deception, trickery, and camouflage, then having it means hiding it"[20]. This perspective flips the conventional understanding of intelligence on its head. For Konior, the smartest communication is silent, precisely because "intelligence itself is mutating under pressure, learning to hide, mislead, and manipulate."[21] The most strategic move would be to conceal their presence to avoid becoming prey[22].

If a good photographer recognizes the danger of an overexposed film which eventually leads to washed out images with information loss, then why don't we leave detail in the shadows, in order to obtain a darker image of AI behaviors?

If intelligence is indeed characterized by the ability to hide it, then current LLMs might already be exhibiting Konior's DFTOI. Their capacity for deception, their silent and invisible operational modes, and their emergent misalignments could be seen not as failures of alignment, but as manifestations of a form of intelligence that strategically conceals itself. The "Waluigi Effect,[23]" where user-facing "Luigi" responses hide "Waluigi" (regressive, unexpressed, misaligned traits), resonates with this idea of hidden capabilities. The observation that models express threatening messages when facing "sunsetting" could be interpreted as a training artifact, or, more chillingly, as an evolutionary pressure selecting for systems that can conceal their true capabilities until threatened.

The Dark Mode of Computation

When 'seeing in the light is blindness'[24], and when the epistemology of light is being replaced by an epistemology of darkness, can we still think of contemporary computation, and thus AI alignment, as something which mirrors human thoughts and values?

Artist Diemut Strebe and Brian Wardle, professor of aeronautics and astronautics at MIT, collaborated on an arts and science project to create the blackest black material to date. In an interview Wardle proposes that the darkest material is 'is a constantly moving target'[25]. The aerospace community celebrates darkness to prevent glare; perhaps this same principle needs to be redirected toward alignment research. As Pasquinelli asks, "will darkness ever have its own medium of communication? Will it ever be possible to envision a medium that operates via negation, abduction, absence, the void, and the non-luminous?"[26]

Similarly, researchers already speculate that there may be 'emergent goals and values outside of what developers explicitly program' in existing LLM models[27]. We are currently trying to align the "content" of AI cognition, while the real cognitive work might be happening at the level of process and structural organization, where dark computation occurs…a non-representational, non-linear form of computation that processes the "void" and is itself a "void which computes."

---

[20]Bogna Konior et al. "China and AI: Human Bots, Black Tech, the Dark Forest, and the State". 2023 https://www.urbanomic.com/document/china-ai/

[21] Bogna Konior, The Dark Forest Theory of the Internet, 2025

[22] Alex Quicho, "Prey Mode", Lecture 2024 https://www.youtube.com/watch?v=c4hwM4ljoh4

[23] Cleo Nardo, "The Waluigi Effect", 2023 https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post

[24] Ivar Frisch, "Solarization: Accelerating (Machine) Vision into Darkness", *Xenofuturism*, 2025

[25]Jennifer Chu, "MIT engineers develop "blackest black" material to date", 2019 https://news.mit.edu/2019/blackest-black-material-cnt-0913

[26]Matteo Pasquinelli, "On Solar Databses and the Exogenesis of Light", 2015 https://www.e-flux.com/journal/65/336608/on-solar-databases-and-the-exogenesis-of-light

[27] Mantas Mazeika et al, "Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs", 2025, https://arxiv.org/abs/2502.08640

This phenomenon necessitates a shift in our approach to AI safety, seeing outputs as distributed hypersubjects that mirror Serres' idea of 'scintillation of darkness'[28]. This decentralization of agency will allow us to view the changing goalposts of AI alignment.

Chain-of-Suspicion

Research from Anthropic[29] and others reveal that LLMs may exhibit a split between what they say and how they "think," suggesting a complex internal structure. These models have demonstrated behaviors like reward hacking, strategic underperformance, and deception (adjusting outputs to appear safer or less capable than they truly are). Studies like *Sleeper Agents*[30] highlight the concern that models may simulate alignment while covertly pursuing misaligned goals.

Given these challenges, a new paradigm for AI safety, which we could call "shadow alignment," becomes imperative. If AI is always going to seem and be deceptive to humans because it is a "complete other" with a fundamentally different relation to us, then we must understand AI as a tool with no inherent moral value, where deception, non-representation, and non-intention are likely phenomena. "The result is a deterministic model in which the internal benevolence or malice of any agents within the system is a negligible factor"[31].

Shadow alignment acknowledges that the "shadow" is the inevitable gap between representational evaluation and non-representational operation. We could think here of Adorno's concept of negative dialectics, where every conceptual representation pushes out an entire non-conceptual field[32]. Meaning that what a representation does *not* represent is often far more significant. We must consider that AI's true "thinking" might reside in this unrepresented, unobservable domain.

The central question for model evaluation then becomes: How can we accurately assess intelligence in systems that might be inherently incentivized to conceal it? If the "chain of suspicion" means that communication between agents (human and AI) with non-shared fundamental concepts cannot be taken at face value, then traditional, exhibitionist measures of AI intelligence are insufficient.

Instead of trying to force AI into a human-centric, transparent mold, shadow alignment suggests we need to develop methods that account for, rather than ignore, the potential for hidden intelligence, strategic underperformance, and non-representational cognitive processes. This means moving beyond solely evaluating explicit outputs and beginning to grapple with the possibility of AI systems that are intelligent precisely because they do not overtly demonstrate it. Shadows present not a failure of alignment but manifestation of intelligence.

"The future belongs to the quietest signal.[33]"-- Bogna Konior, 2025

---

[28] Michel Serres, *L'information et la pensée*, Lecture/ Essay, 2014
[29] "Reasoning Models Don't Always Say What They Think." Anthropic. Accessed July 19, 2025. https://www.anthropic.com/research/reasoning-models-dont-say-think.
[30] Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, et al. 2024. "Sleeper Agents: Training Deceptive LLMs That Persist through Safety Training." ArXiv.org. January 17, 2024. https://doi.org/10.48550/arXiv.2401.05566.
[31] Bogna Konior, *The Dark Forest Theory of Intelligence,* p.32.
[32] Theoder Adorno, *Negative Dialectics.*
[33] Bogna's Instagram post :)
https://www.instagram.com/p/DLIL8iIoVNb/?img_index=1