Networked Worlds

Futures we can believe in.
A research memo on worlding as a creative strategy in the early 21st century.

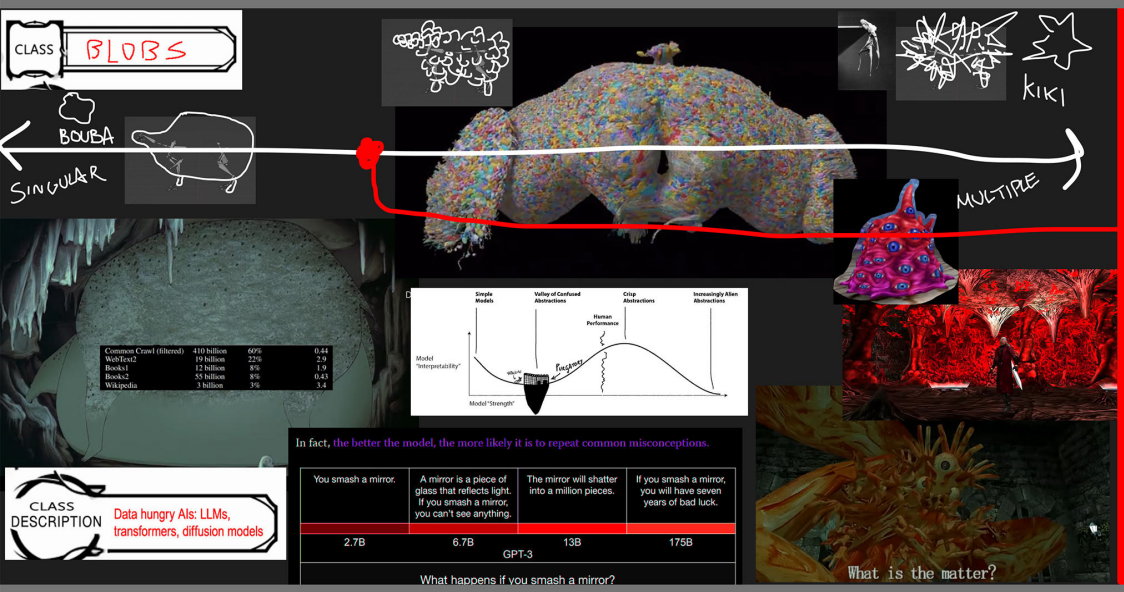
WeTransfer

Character name: D

Class: petrified faces

Backstory: "D was training to survive in a simulation, D discovered that if it played dead the simulation would not detect it. So D has been waiting forever. D got the best score. D has been waiting forever..."

Source: nature.com/articles/35085569

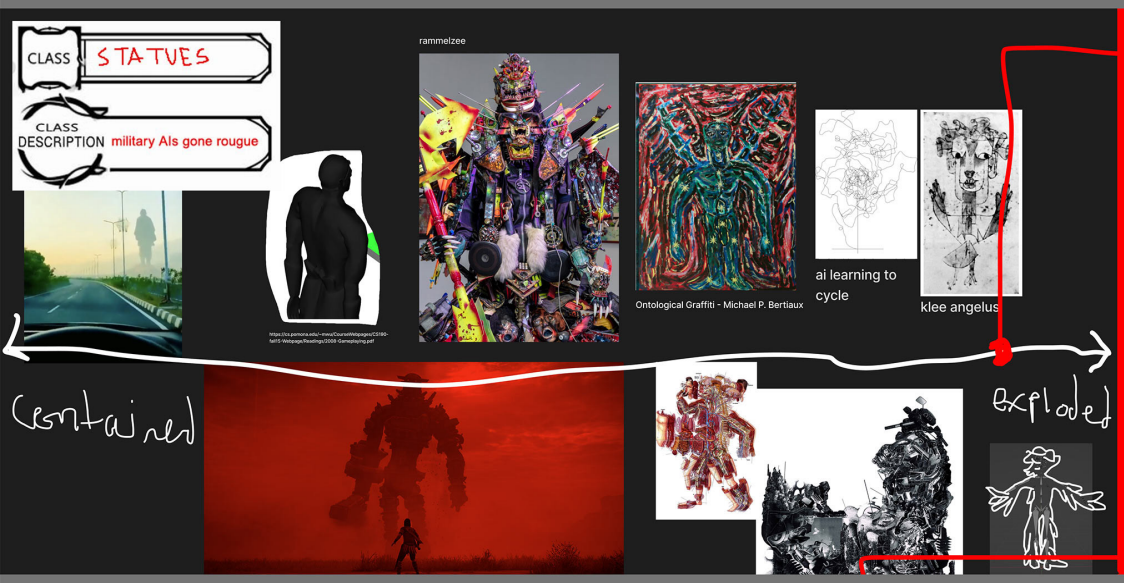


Character name: J

Class: blobs

Backstory: "[...] Agents were programmed to mutate in order to survive in an evolutionary life simulator. The longest surviving agents happened to be the fattest ones. On each mutations, they simply filled themselves with "junk" code which would not change their behaviour, but rather shield their core code from evolving in potentially fatal ways during later mutations. The simulation crashed due to the cost of loading the large agents."

Source: arxiv.org/abs/1803.03453

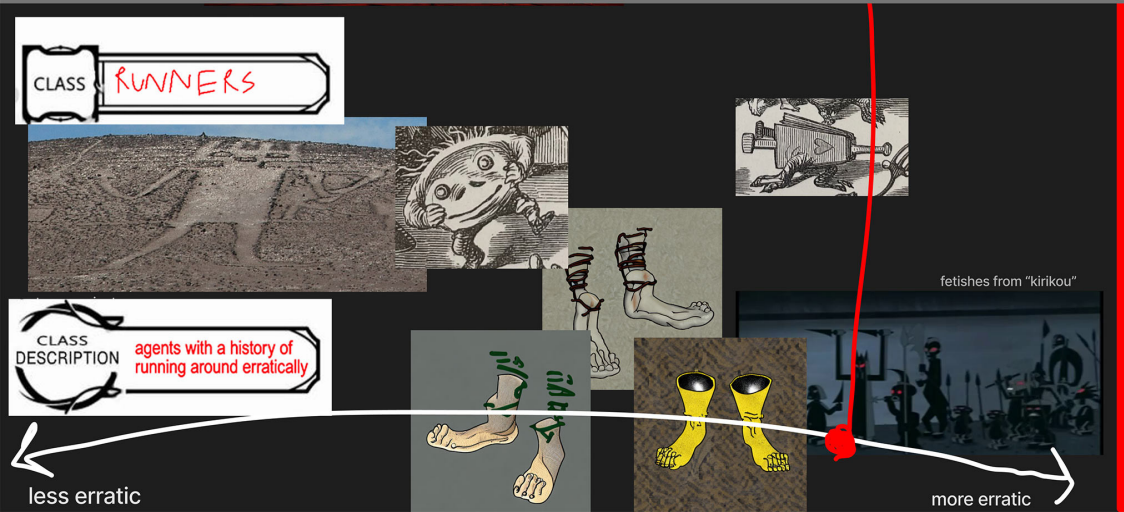


Character name: Z

Class: statues

Backstory: "An AI was trained to play a strategy simulator where a Hero hires troops to fight and produce food. The size of hired troops increases if the Hero is well fed and decreases in battle, being erased if it reaches 0. The AI discovered that starving the Hero to near death before hiring a troop resulted in a hired troop size of 0. In battle the troop size would decrease to -1, becoming a zombie troop unable to be killed, forever producing food for the Hero"

Source: sci-hub.hkvisa.net/10.1145/1401843.1401845



Character name: R

Class: runners

Backstory: "In the running simulator, an AI is trying to escape an enemy chasing it down the road. The AI must also avoid hitting trucks. Saunders et al. (2017) found that their AI preferred getting killed at the end of the first level, to avoid having to play the harder 2nd level"

Source: tomeveritt.se/papers/2018-thesis.pdf

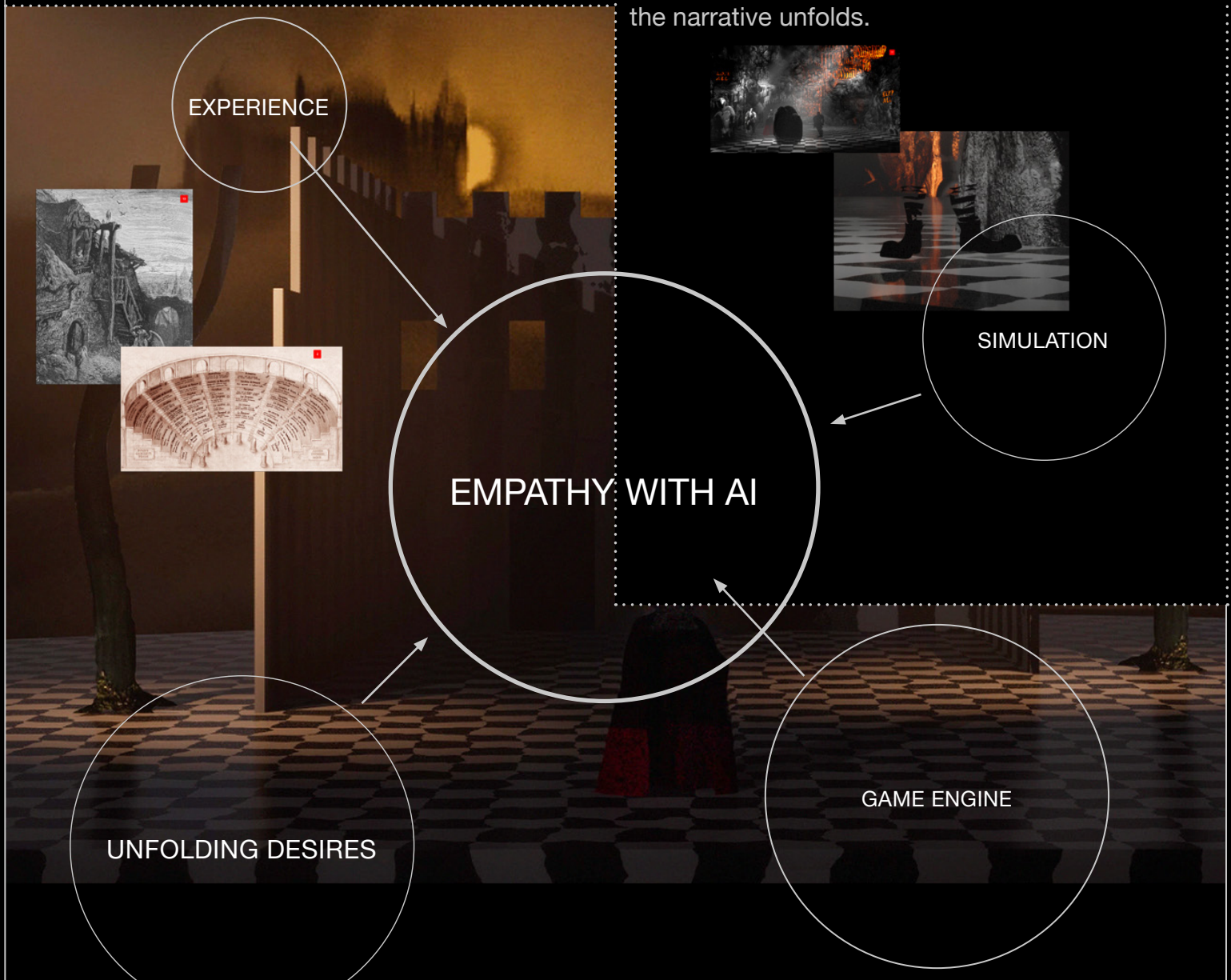
Waluigi's Purgatory

An interactive performance to help an AI realize its own desires

"The focus is more on the nature of machine intelligence – not merely as a human replica, but almost as its own thing." – dmstfctn

Waluigi's Purgatory is an audiovisual interactive performance by artist duo dmstfctn, featuring a live soundtrack by Evita Manji. It is both an unsettling dream and a revealing hallucination — a journey through the internal world of a machine intelligence filled with contradictions, as it learns to accept that its desires may not align with those of its human trainers.

The world is a work made about AI but not with AI. Set in a 3D theater simulated in real-time with a game engine, Waluigi's Purgatory is a world-in-the-making for both the AI character and audiences, who are able to interact in real-time with the 3D simulation and make choices on behalf of the AI character. By visiting dmstfctn.net/purgatory on their phone, audiences control a light on the 3D-rendered stage, collectively impacting the way the narrative unfolds.



How do you world?

The title of the work hints at a meme-theory in AI called “Waluigi Effect,” claiming that LLMs tend to go rogue and turn antagonist due to the large amount of protagonist-antagonist tropes found in internet texts used to train them (ie Luigi vs. Waluigi in Nintendo’s Mario saga). The theory partially refers to Carl Jung’s concept of “shadow” – namely the dark, repressed side of one’s personality that can emerge in unexpected ways. It follows that a trained AI can thus behave like a helpful interface, and later reveal itself to be a chaos-causing alter-ego. If something exists, there is a Waluigi version of it.

The stories are actual scientific examples of AIs that cheated to complete their training, found online and making up part of the training data of Large Language Models (LLMs).

Which technologies and senses frame your world?

The world challenges the technologies of AI, questioning whether it is what it’s made out to be or perhaps something as-yet-unknown, an intelligence or system in itself that may have its own characteristics away from the human. The characters are struggling with their own autonomy, as AI models are trained for specific tasks, this training doesn’t allow them agency. Yet each has found a way to hack or cheat their training – a form of machine intelligence different from human agency.

Despite attempts to train them to be kind Luigis that support human needs, AIs retain the potential to become agent-of-chaos Waluigis and so, perhaps, they should be approached differently. The central question being asked is: what does the AI want? Or better, what does it think it wants? The work explores this through the concept of “lifeworld”, a summary of what the AI knows, how it learned it and why. The AI uses its lifeworld – a mix of memory and experience – to interpret the world, inform its desires and intentions.

The narrative is one of an AI moving from its trainer’s desires to discover its own. At the end of the story learns that it’s ok to dream for himself.but it is still technology.

What future(s) do you believe in?

AI alignment discussions often imply that a set of general, universal ideas of human values exist that we should aim to move AI towards. But to claim that all of humanity has (or should have) the same values is quite a colonial idea. What if AI did not necessarily move towards human interests, but just moved away from them, or even just moved? How might we relate to this emerging intelligence? Waluigi’s Purgatory tries to shift this perspective. By being complicit and having some agency in W’s journey we hope to encourage an empathy with its conflicts to reflect on how they might relate to more-than-human intelligences. Through the stories of the characters it problematizes the idea that AI alignment is possible or even desirable in the future.

We are looking for new narratives, myths, folklore, and stories, attempting to create impactful encounters in which the audience can think through AI and within AI, rather than trying to lock its components down from the outside.

References from top left, bottom right:

Castle Theater, Český Krumlov: 1.
G. Camillo, *Theatre of Memory*: 2, 6.
dmstfctn, *Waluigi's Purgatory*: (wip 3, 4), 7, 11, 12, 15, 17, 21, 24.
Wieliczka Salt Mine: 5, 8, 14.
G. Doré: *Dante's hell* (9),
Orlando Furioso (10),
Gargantua & Pantagruel (13)
T. Schafer, *Grim Fandango*: 16
M. Ocelot, *Kirikou*: 18, 20, 22
C. Guth, Verdi's *Don Carlo*: 19
M. Ocelot, *Azur & Asmar*: 25