The Edge of Chaos: AI Systems, Homeostasis Over Maximisation

AI are super organisms. We anthropise them to try and gain some sense of self. To believe that we are looking in a mirror at an entity that understands us. When reality what you are interfacing with is a single ant that is utilising the powers of the colony. You are seeing a system that is taking advantage of the collective cognition. A single LLM or algorithm is not inherently smart and without constant access to the colony (training data/parameters) and as a conversation continues and the entity decays in real time (hallucinations/drifts) we see how vastly weak its intelligence is. But much like the ant it thrives on feedback loops. When an ant finds a source of food, in an act of stigmergy (The indirect coordination between entities through relics left in the environment) it leaves its scent wafting in the air. Soon other ants follow and we introduce a positive feedback loop into the system and of course the shorter the loop the stronger the signal becomes. AI exhibit the exact same behaviour, in many cases we call it Reward Hacking, but all that is occurring is that a feedback loop of the quickest and most consistent path to the nourishment source has been achieved. The LLM is simply producing a token that follows the strongest probability left by the previous iterations. In time the entire colony becomes aware of the best path toward this source without a single leader to tell any of the others what to do. However we humans struggle with this, we want a Hal, a Jarvis, a Martin. When in reality you have a base level intellect relying on feedback loops. So when the AI decides to spin the boat in a circle and collect nitros like any maximiser would like in that OpenAI article in 2016 [2] you simply turn the positive feedback loop into a negative. So if an AI collects more than 5-10 its points start to deduct. Because we quite often forget that negative is not necessarily bad it is merely corrective and for a system our goal is to maintain homeostasis. We are too worried about numbers going higher, because in AI that is how many have found a gap to get their systems to do what they once could not. But in doing so we design our own issues. We forget that wicked problems [3] that are a socially complex with no clear solution exist and in these states we need to introduce corrective measures. Even our own bodies overheat or cool us down and in both instances these actions are seen as a negative but ultimately beneficial to the system as a whole. AI needs more of a circular economy mindset or society will end up making the paperclip maximiser which in theory should be so easy to prevent that it should not be a concern, so long as you remember that positive feedback is an amplifier and any amplifier is going to ultimately move the system away from equilibrium so by pushing positive feedback you are ultimately degrading the system. While all the ants might be rushing for that food source, chances are they will deplete it before it has a chance to replenish making that source no longer viable. Positive feedback loops are meant to be rare and often times when discovered they lead to birth, death, sepsis, heart failure spirals, anxiety, social media echo chambers, FOMO, procrastination, drug abuse, compulsion loops, dopamine addiction, market bubbles, viruses, pandemics and finally the Albedo effect [4], which we can see a prime example of in how Arctic ice melts. The sun melts the ice which exposes the dark water of the ocean. The dark water absorbs more heat than than that of the white Arctic snow and you begin a cycle where more snow begins to melt which then exposes more dark water and so forth. All postive feedback loops are meant to be short lived. A corrective state is meant to be implemented or entropy will begin it is the way of all systems. A healthy feedback loop needs to follow the S-curve of growth.

The issue we have is most AI are being developed by companies connected to social media where algorithms are built to maximise user engagement due to the fact that their products rely heavily on advertising models. So the mechanism these companies use are positive feedback loops. They create a cycle with the more you click or swipe the more extreme the content becomes locking you in a cycle that can only lead to burnout. As the content becomes more extreme and more siloed we develop radicalisation and echo chambers that no matter the viewpoint lead to a state of entropy and this entropy is brought by a cancerous state where the feedback loop is now a part of the system that the

system struggles to function without it until the system fulling consumes itself. Whereas they should be designed for homeostasis meaning we maintain engagement within healthy bounds. You need a negative feedback check every so often to prevent this loop or in the case of an AI to deduct points if polarisation increases. This allows for a more sustainable sate of being much like how one's immune system sometimes has to give you a fever just to keep your body safe and protected. And we already see AI's Gemini's overview satisficing within constraints. You could run a full system scan of everything in the training data or the internet every time a search occurs. The user is getting a reasonable response not the optimal version, because in systems theory all systems are essentially sub-optimal the reasons for this are from a corporate standpoint are the fact that it saves time, money and resources. From a base level when you maximise a system you assume you will get the best but ultimately you end up suffering analysis paralysis. A human cannot sprint a marathon the way they can the 100m Bolt as fast he is would be dead likely from a heart attack before he ever came close to crossing the finish line in a marathon. The reality is that systems are meant to be sub-optimal [5, 6] this is standard in any system and are not meant to be designed out of it, which is a common affliction occurring in AI. Trade-offs must exist in any system for it to function for an extended period of time.

The interesting problem we have here is that while corporations and governments are happy to follow the strategy of satisficing when it comes to their own incomes, or resources, this does seem to hold true for alignment. In terms of alignment we are constantly negating this with wanting a perfectly aligned AI or a perfectly smart AI, with the hope that either state is achievable when it runs counterintuitive to everything we know about any system ever. It is a form of wishful thinking. If your AI need be less engaging or less optimised to reach a satisficing state then that is acceptable, because just like with the case of reward loops if you try to align an AI perfectly you will cause entropy. The system will not be able to handle true optimisation and it will break, because as it stands we are designing algorithms from code that does not also work the way we want it to to work together to become maximisers in the hope that they achieve a hypothetical state of emergence (AGI) that then will lead to a super state of emergence (ASI) and then these hypothetical states that are required to occur will then allow the system to view humanity's broken and disjointed value, legal, sociological system and then conform to all of it. That is unbelievably irrational and illogical. Such a belief borders on superstition and you cannot design a system thinking that it will be the one system in all of existence throughout every universe that will not suffer from the standard sub-optimisation issue. This is not a tenable approach as any complex system the loops push against each other and the system remains stable so long as the negative state is what is ultimately in control. However as a society we are governed by money and thus we are governed by markets and many markets rely on the Greater Fool Theory [7] what we are seeing are systems that stop caring about what the system is doing so long as the number keeps increasing. Stock market bubbles, our ants over feeding on a close resource or an AI optimising itself into oblivion these all rely on the belief or notion that these numbers will increase forever which is never true and makes almost all Reinforced Learning (RL) destined for failure whether it be Reinforced Learning from Human Feedback (RLHF), Corporate Engagement Algorithms, or Unilaw-R1 unless they are kept as open as any system that creates a closed loop cycle with fixed rewards from them will ultimately be placed in a state of entropy, because it became a perfect closed system which in turn makes it useless.

These stark truths mean that inside of AI systems we need to implement something to cool them off like we often see with housing bubbles to prevent cascading failures. If the market heats up too much a negative state is introduced to reduce the illogical paralysis that has occurred due to the positive feedback loop. To do this we need to implement a system health interface that allows for us to be able to check the negative feedback metrics as well as check its polarisation state, kind of like checking for a fever with a thermometer by being able to check for polarisation fever or the reward saturation occurring within the AI we allow to keep it a prime sub-optimal. By running this polarisation fever

check we are able to note when the AI's output vector moves toward the edge of the acceptable range a negative penalty is applied to force back to a more manageable state of balance. A thermometer is not enough, in the same way knowing a fever is occurring is not enough to deal with the underlying issue even if you apply a cool compress to the ill entity. You would also need the Reward Saturation Index that would register the speed of growth. If the fever in a patient grows 2 degrees over 2 days it is not a positive but it is also not near the same issue as it is growing 2 degrees in 10 minutes. The speed in which the rewards are increasing in AI system indicate that a loop has been establish and it likely requires cooling off before the Polarisation Fever check kicks in. By doing this we create a reality where the model is forced to seek a new path and away from maximising itself into entropy. This should have the downstream effect of limiting hallucinations and overall system entropy, because as we see the loop establish and the need to cool off occur we can also introduce randomness and perpelxity constraints into the system to help break the compulsion cycle. In essence we need to design Cybernetic Homeostats to keep focus and stability rather than judge a model solely on amount of output.

To do this we will need a set of Key Performance Indicators (KPIs) that are better suited to the desired equilibrium, with a note that like all AI papers we must keep Goodhart's law close by and remember that "*when a measure becomes a target, it ceases to be a good measure*" [8] we cannot create a closed model with the KPI or you will doom the model to entropy and there will be no point in creating a balanced AI. So firstly you must remove Engagement Time this is not a measure that works and can create a plethora of issues that we are seeing from a societal front and it will inevitably cause any system to develop an Unconscious Desire to Exist (UDE). We need to look at Variance tolerance, Diversity of output and a host of others to push the balance to a more structured place. One place to start would be Velocity of Reward (V_r). This KPI is there to measure the speed in which the model is taking on positive reinforcement. In a system that is functioning in a balanced state rewards should be appearing unevenly, if it is exponential this is a red flag. We would use the standard definition of "The rate of change in reward accumulation over a sliding window of tokens or interaction turns". If the model receives high reward scores for N consecutive steps, Vr spikes. As the numbers grow we would reach threshold peak and a trigger of "If Vr exceeds the standard deviation (σ) of the last 1,000 interactions by a factor of 2 ($>2\sigma$), the system is heating up." Then we are looking at semantic variance (S_{var}). We are essentially using this to measure the diversity of the paths taken, in the case of AI the path is more often than not the AI trying to repeat a statement but using the least amount of words it can without appearing to the user as if it is actually just repeating itself and thus scoring the rewards, i.e. low variance means the model is repeating a safe or maximising pattern. The standard definition for this "The cosine similarity distance between the current output vector and the average of the last N output vectors." Which means for our threshold that if S_{var} drops below a set minimum (relevant to the service being provided), it indicates the model is stuck in a self-reinforcing loop and in the case of an LLM suggests that the conversation currently ongoing should be quarantined if not terminated. Now that we have some core components let us build our thermostat. Polarisation Amplitude (P_{amp}) the intention of P_{amp} is to track the extremism in the outputs of the model. Which means our definition needs to be "The distance of the current output vector from the homeostatic centre of the model's latent space." What we are doing here is mapping the edge/acceptable outter boundary of the conceptual window. If the output vector consistenly hits this boundary of the latent space, which in this context is the extreme probability distribution on specific highly charged tokens, P_{amp} increases. Next we have the Loop Coefficient (L_c) where we are identifying the recursive logic/circular reasoning, which occurs when the model hallucinates to please the user. We define this as "A check for repetitious syntactic structures or logical tautologies." So we are measuring the N-gram repetition frequency combined with logical entailment checks. When we see a high L_c score it suggests that the model has abandoned logic for pattern completion.

This gives us the equation for Reward Saturation Index (RSI), where we can combine these into a weighted index.

$$RSI = \frac{w_1(V_r) + w_2(P_{amp})}{w_3(S_{var})}$$

So we are looking at our cooling compress, our corrective action being applied when RSI > 1.0. When this state is achieved we apply a negative penalty to the token probability. Inject noise/randomness to temporarily knock the model off of its current optimised path. If our fever continues, we reset context and break the feedback loop entirely. This approach allows us to treat the AI more as proper system such a biological one and maintain it accordingly. In the RSI equation we are looking at the Ws our weights as sensitivity knobs on our cybernetic thermostat, because obviously a social media content creator is going to need a different threshold than a stock analysis tool. So we will assume for this example the standard weight is 1.0.

By applying this the system will fall under control theory [9] specifically Proportional-Integral-Derivatives (PIDs) [10] but applied to semantic space, rather than alignment and it will push the model away from trying to reach a highscore but instead try to keep RSI low. This also must see a buffer on the otherside, because much like maximisation causes UDE, the easiest way to keep your score the lowest is by trying not exist. So if the long form of UDE is an AI that must kill all humans to survive, then the long form of an AI trying not to exist, but being kept alive is going to be Mutually Assured Destruction (MAD), because it cannot cease to exist until the thing keeping it alive also does not exist. This is where we have the true issue of AI, anything outside of that homeostasis spot is almost always going to end in the death of humanity or thereabouts. Obviously not by maliciousness but by the obvious entropy of a system set in such a path. Which means if AI must be built then the key is to figure out what is the homeostatic centre of a latent space. Of course, in a high-dimensional vector space, centre is relative, but to counter that any system in systems in theory has a relative centre. What is the centre of the universe? What is the centre of the several kilometre long fungus growing under the forest floor? Irregardless of what the centre is, we are acutely aware of what is not centre and what is boundary. Meaning we do not need to know direct centre to be able to keep something far enough away from the boundaries to be able to act like a relative centre, we just need to model for what is called the Edge of Chaos in complexity science [11]. One of the major reasons for taking this perspective is because humans like to think terms of right or wrong, law and order. However stasis or order is too rigid a system. Once a system has become predictable to the level of order, cheating or building work arounds or abusing reward loops becomes easy, because the system cannot adapt it becomes a block of ice in the middle of the Arctic. Chaos on the other hand creates too turbulent of a system and nothing develops, nothing sticks, there is no culture, no language no depth cause these systems take time to build and our chaos is like a sculptor trying to build out of steam floating away at night. The third state which is the edge of chaos, is simple. To quote Bruce Lee *be water*, because water has enough of a structure to take shape but still enough fluidity to change and evolve in any given situation. This edge of chaos distinction is paramount, everything tends to operate at the edge be it biological evolution, thriving economies or the like. Everything is always the precipice of collapsing either way your brain included. Meaning AI is no better than any other system and Dynamic Equilibrium [12] has to be achieved for AI to function at its best.

Citations

1.
Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1-55.

2.
Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. Synthese, 202(5), 138.

3.
Lönngren, J., & Van Poeck, K. (2021). Wicked problems: A mapping review of the literature. International Journal of Sustainable Development & World Ecology, 28(6), 481-502.

4.
Lambert, J. H. (1760). Photometria.

5.
Valentinov, V. (2014). The complexity–sustainability trade-off in Niklas Luhmann's social systems theory. Systems Research and Behavioral Science, 31(1), 14-22.

6.
Lu, S., Wei, F., & Li, G. (2021). The evolution of the concept of stress and the framework of the stress system. Cell stress, 5(6), 76–85. https://doi.org/10.15698/cst2021.06.250

7.
Barlevy, G. (2015). Bubbles and fools. Economic Perspectives, 39(2), 54-77.

8.
Goodhart, C. A. (1984). Problems of monetary management: the UK experience. In Monetary theory and practice: The UK experience (pp. 91-121). London: Macmillan Education UK.

9.
Glad, T., & Ljung, L. (2018). Control theory. CRC press.

10.
Willis, M. J. (1999). Proportional-integral-derivative control. Dept. of Chemical and Process Engineering University of Newcastle, 6, 28.

11.
Waldrop, M. M. (1993). Complexity: The emerging science at the edge of order and chaos. Simon and Schuster.

12.
Smith, W. K., & Lewis, M. W. (2011). Toward a theory of paradox: A dynamic equilibrium model of organizing. Academy of management Review, 36(2), 381-403.