All trials are unethical: understanding as ‘entropy descent’ and the judicial consequences.

Introduction:

We have been working to remove perceived systematic errors from our judicial apparatus for a very long time, but the model and its output —the verdict— has no way to be evaluated. In machine learning terms, there is no fitness function, no carefully verified data to test the model on. This near complete lack of fine-tunability and a quite unknowable accuracy exposes the system to random errors. We are, of course, aware of the possibility of these errors and use a multiplicity of jurors and judges to nullify them; however, we are treating the symptoms not the system, plastering prejudices and excising subjectivity rather than getting at their origin. We will here diagnose the problem, putting a name to it and encouraging awareness such that juries can go forth with intimate knowledge of the mechanism behind their mental deliberation.

The Algorithm of Understanding:

As the brain traces the abstraction hierarchy upwards and causality from proximate to ultimate, a measure called dynamic signal entropy decreases (Mago et al., 2026). This tracks the future possible trajectories and evolution of a state. I am increasingly convinced that this is not a byproduct but the intended purpose, that minimisation of dynamic signal entropy (from now on just “entropy”) along both its own dimension and the temporal one —i.e., finding the shortest chain/earliest possible minimum— evolved for energy-saving optimality. I imagine percepts starting on some turntable which feeds into slides (abstraction/causal chains) at which point the process of understanding is finding the slide that goes lowest to minimise entropy—‘entropy descent’. The initial neural dynamics are thus the most opportunistic, having a wider space of future evolution, and abstraction/progressive association works to prune both the nonsensical and unlikely future trajectories and the number of futures themselves via top-down flow of constraints, funnelling dynamics into a lowest entropy —and not necessarily valid— interpretation of the percept(s). This may act stepwise with any one node (neuron/column/circuit etc.) taking the node with the least possible futures as its next step, but it definitely acts end-wise, searching in general for a node with no futures —and it is my posit that we look for the closest null-potentiality node. This makes complete evolutionary sense: instead of looking for the full abstraction trace (or a maximised one) which early humans would never have been able to find, it is instead much less costly to find a shorter chain ending with an isolated abstraction which survived constraints and is for some reason acceptable to the brain; indeed, isolated in the thermodynamic system sense, in that there will be no further energy transfer. In each attempt at understanding, our brains search for the monolithic neuron, the lonesome cycle or circuit of columns connected to no other, the end of energy transfer and the descent of entropy.

Validation-wise, this model is reminiscent of Levinthal's paradox of protein folding which is solved similarly with a "folding funnel" which chains travel down as they become more energetically stable and less inherently entropic as shape potentiality decreases.

We can interpret this want for minimum entropy in many ways: firstly, as want for maximum predictability (minimum possibilities) which we can ground in our symmetrical concept of beauty, our enjoyment of patterned music and narratives, and the neuronal least-action principle. Secondly, as a least-further-action principle with the brain looking for the specific chain of thought that is, for some reason, satisfactory, satiating curiosity. Thirdly, and the most important interpretation, as maximisation of local explainability: if we construe future trajectories as questionability, then minimisation of this is maximisation of explainability. We can further validate with the converse: greater fear arises when there is greater entropy (less explainability) because increased uncertainty and wider possibility give the opposite response to the energetic isolation of understanding (repeated bursts through all the chains with no resolution) much like in psychological horrors where we seldom see the monster and so there exists a great many further abstraction chains diverging from its silhouette; this is what fear of the unknown is, the existence of multitudinous futures. We look, then, for the first thing that stops the questions in the thought process, the earliest collapse of the space of thought chains.

Our brains thus do not seek to trace the full length of abstraction or causal chains but minimise future branching, alternate explanations, and further questioning. This would explain the opportunistic appearance and appeal of supernaturalism like God(s) and simulation theories, manifesting as higher level forms of agenticity. Indeed, the key here is that understanding evolved for energy-saving optimality not cognitive optimality.

This principle of earliest explainability explains all prejudices: because previous personal observations have impressed a commonality (and so a causality), a top-down flow of this constraint closes the mind to other possibilities leaving stereotypic explainability as the earliest entropy minimum, employing preconceived causality rather than analysing a specific causal chain. Why go looking for lengthy and well-resolved causal chains when the brain is prepared to just accept a subjectively statistical fact?

If our brains default to states that maximise the internal world model's explainability, which is unknowably calibrated, then how can we rely on any judgments? Entropy descent is a homogenous algorithm generating heterogeneous understanding biased towards that which is not questioned by that brain. And so, if the brain defers to individualised abstraction cul-de-sacs instead of travelling down the full motorway, we can never reliably reach the intended destination—the truth. Additionally, when there are multiple accounts or suspects, the brain, to be unbiased, would have to hold each causal or event chain in superposition, leading to a sort of zero-point entropy which is in direct opposition to the normal-functioning of the brain.

There are also questions of prosecutorial policy as to which cause to go after: entropy descent towards earliest explainability very literally translates to a preference for proximate causes which means we often miss out on the ultimate causes in our understanding. Indeed, those who orchestrate crimes are harder to pin down, let alone prosecute, than those who commit them. Most countries have laws with provisions for both conspiracy to commit and actual committal, but conspiracy and its rather intangible nature are far harder to prove legally and mentally; in the expanse of the causal network it is rather like blaming butterflies in Brazil for hurricane Katrina. It is this chaotic causality and lack of natural, readily available explainability that led us to artificially explain weather with deities —this should not offend as it is the frequency of theism which we are picking out, not any specific one. This capacity for artificial explainability is of course another ethical concern: when there are not any available minima, jurors may well create one, alternatively, when there are several with the same dimensions, they may arbitrarily pick one.

From a quantitative angle, the No Free Lunches theorem (NFL) implies, with great unethicality, that we cannot have an optimal judicial algorithm for each situation (assuming, of course, a wide enough space of testimony). Indeed, it may be helpful —and certainly illustrative— if we consider NFL to be a Heisenberg-esque principle in which the product of algorithmic narrowness (performance in a specific problem) and algorithmic generality (performance across problems) are always less than or equal to a certain resource-dependent constant, then, in general:

$$N(A,p) \bullet G(A,P) \leq C(R) \;\; N,G,C > 0$$

where N is the narrowness, G is the generality, A is an algorithm, p is a specific problem, P is the problem set/space, and R is the resources (compute, data volume, inductive and inferential biases, training time, etc.). This tradeoff is a fact, if you increase narrow performance, you decrease general performance. Entropy descent seems to maximise generality—whilst we are generally good at finding situational understanding, we are not good at finding a specific situational understanding. We are thus unlikely to achieve the narrowness of biases and assumptions sometimes required to understand the truth; visually, we can rarely walk the tightrope without a fatal step towards convenience.

Conclusion:

A verdict is the adversarial output from a selection of guilt models; this multiplicity and active selection processes —jury selection and judgeship confirmation— have removed overtly biased entropy minima. Yet the tendency to settle in minima, the earliest, easiest explainers, remains in all models. This is of great importance: we might think we have stripped the apparatus of errors but we have only removed that which the majority can

see, not that which afflicts the majority. We have presented thus not a new ethical issue but a new domain of an older one.

Truth is so often stranger than fiction, yet we are all predisposed to subsume the most self-sensical explanations, subconsciously closed to more untoward accounts, inconsiderate of more convoluted chains, and ignorant to the more energy and entropy intensive truths. It is surely our ethical responsibility to make jurors conscious of this, just as informed consent advises the truth of a patient's medical prospects, informed deliberation ought advise jurors of the state of their mental faculties.

Reference:

Mago, J., Chandaria, S., Miller, M., & Laukkonen, R. (2026). Pure awareness, entropy, and the foundation of perception. *Neurocomputing*, *684*, 133443. https://doi.org/10.1016/j.neucom.2026.133443