Symposia of the 2025 convention of the UK Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB), held at UWE, Bristol.

Keynotes

Models of Life: robots as scientific instruments

Prof. Alan Winfield
University of the West of England

I outline a new book I am writing called Models of Life. The book is about robots that have been designed primarily as working models to test hypotheses about (aspects of) animal behaviour, intelligence, evolution and, in humans, morality, culture and consciousness. Each of these core chapters includes a number of case studies which explain how the robot models work and what they teach us. I set myself fairly strict criteria for selecting these case studies. Bio mimicry is not enough to warrant inclusion. There must also be hypothesis testing of value to biologists, or (in the later chapters) moral philosophers, ethnologists and philosophers of mind. The talk is illustrated with brief outlines of some of the case studies, starting with Grey Walter's machina speculatrix.

Artificial Intelligence and the Simulation of Consciousness

Prof Owen Holland
Sussex Centre for Consciousness Science

In 2006 I had given an AISB plenary entitled "Artificial Consciousness and the Simulation of Behaviour" which described the ongoing work on the development of what was intended to be a conscious robot, CRONOS. Nineteen years later the invitation to give another AISB plenary seemed to be an excellent opportunity both to review the project and to see where the train of thought had led. I was able to use parts of the original presentation to describe the technical robotic progress, which gave rise to further progress within robotics, but we had not really come within touching distance of consciousness itself because of the poorly structured will-o-the-wisp of the concept itself.

How to proceed? Rather than using a logically designed step by step research strategy, I proposed a reliance on emergence. I reviewed cases where the unexpected emergence of a phenomenon had solved the problem at hand and described three instances of the role of weak emergence in my earlier work at UWE. There is a widespread belief that consciousness Is a case of strong emergence, but as this is an intuition rather than a formally derived position, I argue that emergence may well be a path worth pursuing.

The CRONOS project centred around the development and exploitation of internal models of the body/self and the world, and I argued that the remarkable neurodiverse gymnast Simone Biles could be an example of the existence of a close connection between her model of her body and her consciousness.

However, the major difficulty flagged up by the CRONOS project was the requirement that consciousness should be embodied in a physical robot. There is a slight but perceptible drift in the direction of finessing the sheer trouble and expense of using a physical robot by exploring the possibility of creating a conscious virtual robot – possibly a digital twin of a physical robot – and the philosophical groundwork has begun, and the topic is still both novel and challenging. A possible stroke of luck recently occurred in the development of the concept of the metaverse – a virtual world as complex and detailed as the real world within which a virtual entity might be influenced in the direction of developing a form of virtual consciousness – and although the metaverse initiative seems to be flagging as it has failed to find a path to significant monetization, its potential may eventually be realized in the very different area of artificial consciousness.

General Track / Cybernetics & Robotics

Cybernetics is a multidisciplinary field that studies systems of communication, control, and feedback in both living organisms and machines. It is primarily concerned with how systems regulate themselves, maintain stability, and adapt to change using feedback loops. Cybernetics has been applied across many domains, including biology, engineering, and social systems, and its foundational concepts have significantly influenced artificial intelligence, and robotics. Cybernetics provides the theoretical framework for designing robots that can process information, learn from their environment, and adapt their behaviour. Robotics, in turn, applies these principles to create machines capable of performing complex tasks with varying degrees of autonomy, often mimicking biological systems.

Steve Battle (Conference & track Organiser)

Where ChatGPT and other LLM-based meaning-making chatbots fall short – and where they succeed

Joel Parthermore
Department of Cognitive Neuroscience and Philosophy, University of Skövde

keywords: meaning making, large-language models, breakdowns, Imitation Game

Abstract

A prominent philosopher of cognitive science recently asked on a popular social-media site "why is the success of AI not impeded by the inefficacy of meaning?"

The answer turns on what one means by "success". What is so striking about LLMs is not their relative conversational fluency (as impressive as that is) but just how quickly and easily, with the right prompts, that fluency breaks down. (The presentation will offer examples using ChatGPT4.) The lesson they offer is just how far the mere semblance of meaning goes. (That applies, too, to everyday human interactions.) Nevertheless, there is no reason to think that the mere semblance of

meaning can be sustained for any period of time without the active (if, perhaps, unreflectively aware) participation of the other (presumably human) partner. The breakdowns are most obvious where one exploits the systems' lack of self-reflective capacity, and it's not clear how that can be corrected without actually giving the systems the self-reflection currently limited to a select set of living species. (Some of the strongest proponents of AGI-through-LLMs seem to believe that that self-reflection will come for free.) Many of the fixes that help these systems escape traps now are kludges of little more sophistication than Eliza showed more than 50 years ago. Noam Chomsky is right: although artefacts may, indeed, be meaning makers in the human or non-human animal sense, there is good reason to believe that LLM technology is not the path there. Lacking life (and no one is claiming that these systems are alive or about to come alive), LLMs are reactive rather than

anticipatory prediction machines. They have no defence against their capacity to "hallucinate" meaning, putting the burden on the human agent in the loop to judge the reliability of the output and assign meaning to it. They cannot create their own meaning, and sooner or later, that creates problems.

It is revealing that, so far as I can tell -and despite claims of ChatGPT "blowing past" the Turing test -no system has been able to succeed at the Imitation Game (Turing's own name for the test) according to the original rules -which included both players who are trying to convince the judge being privy to each other's conversations, along with the possibility (for slow typists!) of passing messages verbally through a third party and encouragement to the judge to ask trick questions

inspired by their knowledge of the artefactual agent's likely weaknesses. That's despite Turing's prediction that a system would be able to succeed modestly well within 50 years of his writing. It's been 75 years since Turing drafted the paper, and artefactual systems still can't do it!

The conclusion of this paper is that LLM-based systems, as AI, are harbingers of yet another AI winter. They are staggeringly energy hungry at the same time they can offer only a pale shadow of a living agent's self-reflective (far more energy efficient!) meaning making due to limitations intrinsic to their nature. They risk being successful in another way, however: as expert and tireless producers of misinformation at a time when the traditional news media are struggling to survive.

This paper is a companion piece to a presentation I made at the AISB convention in Swansea in 2023.

Neural Analogue PID Control

Steve Battle

School of Computing and Creative Tech., University of the West of England, Bristol.

Keywords: PID, neuro-control

Abstract

Neural nets have been used to implement Proportional Integral Derivative Control (PID) control systems, but these are typically defined using clocked artificial neurons, with access to the difference between inputs at discrete times t and t-1. These exact time intervals are not available to analogue computers, nor indeed, biological brains. Continuous Time Recurrent Neural Networks (CTRNNs) are neural nets that function as dynamical systems interacting with an analogue world, defined as a system of differential equations. As this model is biologically inspired, we can look for comparable neural circuits in biological systems.

A single neuron in the input layer calculates the error between a reference signal and input from the controlled variable. A middle layer contains three neural structures comprising P-circuits, I-circuits, and D-circuits. The P-circuit is an artificial neuron with a single error input where the output is proportional to the input averaged over a time constant. It forms an exponential moving average. The I-circuit is a recurrent excitatory loop that integrates the error over time. Discrete time models have the luxury of feeding back and subtracting output from t-1. In continuous time models this difference is vanishingly small for infinitesimal time steps. The D-circuit forms the difference between the error input and the output from an integrator; a technique inspired by analogue computing. An output layer forms a weighted sum of the P, I, D circuits, which is then output to the controlled plant. The PID CTRNN can be used for real and simulated robot control, for example, to maintain a constant speed over uneven ground.

1. Introduction

Proportional-Integral-Derivative (PID) control is a widely used negative feedback control system in industry and indeed shipping where it is literally used for steersmanship; the origin of the term ‘cybernetics’, deriving from the Greek for ‘steersman’ (kybernetes). A PID controller continuously adjusts a process variable, like heading, to match a desired setpoint by determining the error from the setpoint and then calculating the proportional, integral, and derivative contributions to the control action. The classic Watt governor of 1798 is a proportional controller that maintains the speed of rotation of a steam engine as measured by a pair of centrifugal pendulums, using this to adjust the steam inlet valve in proportion to the error (Mayr, 1970). The PID controller was developed by Nicolas Minorsky (1922) for the automatic steering gear of the battleship New Mexico, which underwent trials for the US Navy in 1923 (Bennet, 1984). The integral component reacts to the accumulation of past errors and can accommodate factors such as the wind and tide. The derivative component can respond rapidly to change, and even anticipate future error based on the rate of change. Despite the success of these sea trials, the new PID controller was removed from the New Mexico on the orders of Captain H.S. Howard who explained that “the operating personnel at sea were very definitely and strenuously opposed to automatic steering, and they wished us to have nothing further to do with it after these tests were completed.”

2. The Gamma Loop

In *Cybernetics: Or Control and Communication in the Animal and the Machine,* Norbert Wiener (1948) explores the commonalities between biological organisms and machines that regulate their behaviour using feedback loops. One biological feedback system that has been investigated is the Gamma loop (Mihailoff & Haines, 2015) for muscle control named after the alpha and gamma motor neurons responsible for the control of muscle activation. According to Powers (1981), the gamma loop can be understood as a proportional feedback control system. Skeletal muscles connected to your

bones comprise thousands of separate muscle fibres that generate a contraction force. Muscles are arranged antagonistically so that their contraction forces pull in opposite directions. These muscles are activated by alpha motor neurons. However, muscle tissue provides no direct sensory feedback. Instead, sensory information about how the muscle is being stretched is provided by *muscle spindles* embedded alongside muscle fibres. Muscle spindles are not just embedded sensors but are also muscles themselves, activated by gamma motor neurons. The musculature doesn't provide any significant force but keeps the muscle spindle contracted, as the spindle produces an error signal that is proportional to the stretch only when the spindle is under tension. The alpha and gamma neurons are co-activated by descending nerve fibres from the brain so that when the muscle contracts, the muscle spindles also contract, maintaining tension. The control loop is closed by (*Ia*) sensory fibres that re-excite the alpha motor neurones. When the muscle is involuntarily stretched, this additional excitation acts to counteract it by exerting more force, in proportion to the error (Purves, 2008).

Powers (1981, p.93) left us with some homework, stating that "*though this linear analysis is certainly not a wholly accurate picture, it will serve for a while."* Is there evidence for biological control systems using integral control? Integration appears to occur within motor neurons. At low frequencies of stimulation, each action potential in the motor neuron results in a single twitch of the related muscle fibres. At higher frequencies the twitches sum to produce a force greater than that produced by single twitches. Integration appears to be a natural function of the motor system.

And what of derivative control? Motor units are not uniform but vary in size according to the number of muscle fibres that are innervated. Small motor units are activated by weak synaptic stimulation and produce rapid muscle movement for fine control. Larger motor units have a higher activation threshold and react more slowly but provide more force. According to the size principle of motor neuron recruitment, smaller motor neurons are activated first, followed by larger motor units. The small motor units may act as high pass filters, providing rapid but small corrections in response to small perturbations in the error signal. These high-pass filters can be seen as differentiators. In many signal processing applications, high-pass filters are used to perform differentiation for signals with a frequency below the filter's cutoff frequency. Additionally, within the muscle spindles there are two different types of sensory *intrafusal fibres*. *Static nuclear bag fibres* are sensitive to changes in muscle length and are ideal for proportional control, whereas *dynamic* nuclear bag fibres are mainly sensitive to the rate of change in muscle length. This differential signal is superimposed on the same (*Ia)* sensory fibre that carries the proportional signal as before, and this may serve to emphasise high frequency disturbances.

3. Neural PID Control

If we accept that PID control is biologically plausible, then how might we implement this within a neural network? Shu and Pi (2000) describe a discrete time neural network of this kind called PIDNN. Both the Integral and Derivative activation functions are expressed as functions of time t (the present) and time t-1 (some point in the past). Think of these Discrete-time recurrent neural networks (Forcada, 2002) as clocked systems where neural activation is recalculated at regular intervals. While this approach provides excellent performance, it is not consistent with unclocked biological systems where only information from the instantaneous present is available. An *analogue* approach to modelling neural nets uses only current values and their rate of change.

4. Continuous-Time Recurrent Neural Network

We propose a Continuous-Time Recurrent Neural Network (CTRNN) model of PID control. CTRNNs are neural nets that function as dynamical systems interacting with an analogue environment. They are defined by a system of differential equations.

The action potential of each CTRNN neuron is modelled as the rate of change in the weighted sum of its inputs as in Equation 1. A time constant, $\boldsymbol{\tau}$, controls the response rate of the neuron. The activation of the neuron is a

function, g, of this potential minus a bias which may be 0. A simple saturating linear activation function will suffice here, that applies a sharp cut-off to values outside the range [-1, +1].

$$\tau \frac{dx}{dt} = \sum_i w_i y_i - x$$

$$y = g(x - \theta)$$

where

$\boldsymbol{\tau}$: time constant

$\boldsymbol{\omega}$: *weight*

$\boldsymbol{x}$: membrane potential

$\boldsymbol{y}$: activation

$\boldsymbol{\theta}$: *bias*

Equation 1: CTRNN neuron model

Figure 1 shows the architecture of the neural PID controller as a small three-layer artificial neural network. In the figures, hollow synapses represent excitatory connections, while solid synapses indicate inhibitory ones. The input layer is a comparator that receives inputs from the (controlled) process variable Q and the desired set-point (reference). The reference is an excitatory input while Q is inhibitory, so the combined effect is that the output of this summation neuron is the difference, an error signal. The middle layer has three independent sub-circuits that compute the Proportional, Integral, and Derivative contributions. The details of the P, I, D circuits are not shown here and are developed later in this paper. The output layer calculates a weighted sum of the PID contributions to form the control output. This output and disturbances from the environment are summed into Q. The goal of the PID controller is to counteract these disturbances.

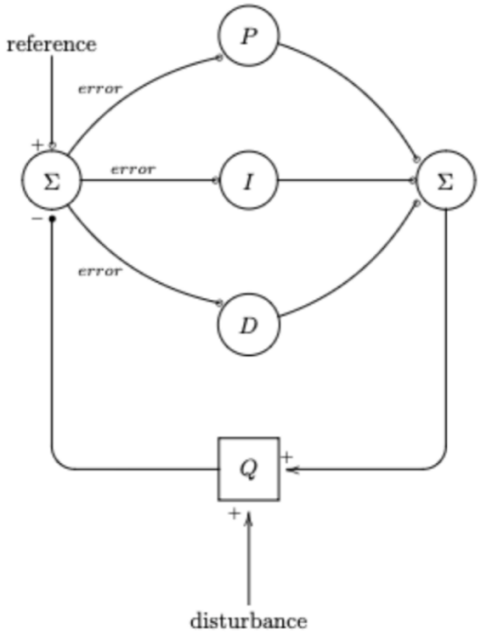

Figure 1: Three-layer neural PID controller

5. Proportional P-circuit

In a P-circuit (proportional control), the output is proportional to the input averaged over a time constant as defined by Equation 2. The input is the error computed by the comparator as the difference between the reference and the controlled variable. The P-circuit may be used by itself in a proportional controller as in Figure 2. In most applications of proportional controllers, the output signal is incremental. For example, if the controlled variable is a ship's heading and the output controls the angular position of a rudder then once set and the error is eliminated the proportional output drops to 0, disturbances from wind and tide aside. This is not the case with motor neurons which must maintain (at the level of a neural population) a continuous level of activation to produce a given muscular force. A simple proportional controller would oscillate wildly in this scenario.

$$\tau \frac{\delta P}{\delta t} = \varpi_i \, error - P$$

$$y_P = g(P)$$

Equation 2: P-circuit output is proportional to the error.

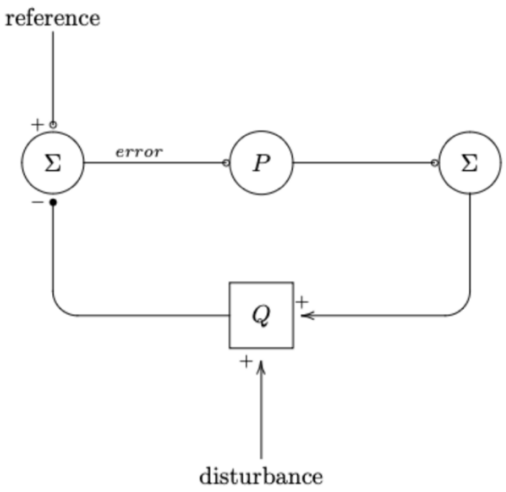

Figure 2: Neural proportional controller

6. Integral I-circuit

The I-circuit functions as the system's memory, employing a recurrent excitatory loop to integrate the error over time as defined by Equation 3. In a PI controller as seen in Figure 3, the proportional and integral outputs are summed to produce the control output. This approach works well where the output is not incremental, and a given output needs to be maintained. The I-circuit produces a baseline output, while the P-circuit handles disturbances on a shorter timescale.

$$\tau \frac{\delta I}{\delta t} = \varpi_i \, error + \varpi_j y_I - I$$

$$y_I = g(I)$$

Equation 3: I-circuit as a recurrent excitatory loop.

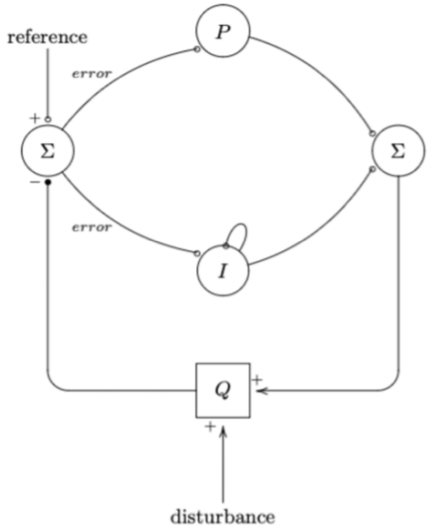

Figure 3: PI neural controller

7. Derivative D-circuit

The function of the D-circuit is to approximate the instantaneous rate of change of the error signal. As noted earlier, in discrete-time models, it is straightforward to subtract the output from the previous time step, to compute the difference between the input at time t and at t-1. In continuous time models the difference becomes infinitesimal. Using an approach inspired by analogue computing we incorporate an integrator into the loop, to act as a memory operating over a given timescale. The D-circuit computes the difference between the error input and the output from this integrator. The equations defining the difference-integrator pair are given in Equation 4, below.

$$\tau \frac{\delta D}{\delta t} = \varpi_i \, error - \varpi_j y_\iota - D$$

$$y_D = g(D)$$

$$\tau \frac{\delta \iota}{\delta t} = \varpi_i y_D + \varpi_j y_\iota - \iota$$

$$y_\iota = g(\iota)$$

Equation 4: D-circuit difference-integrator pair.

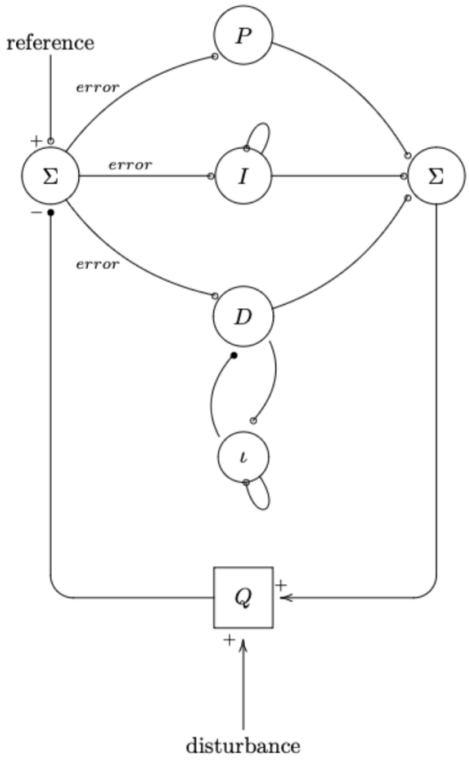

Figure 4: PID neural controller

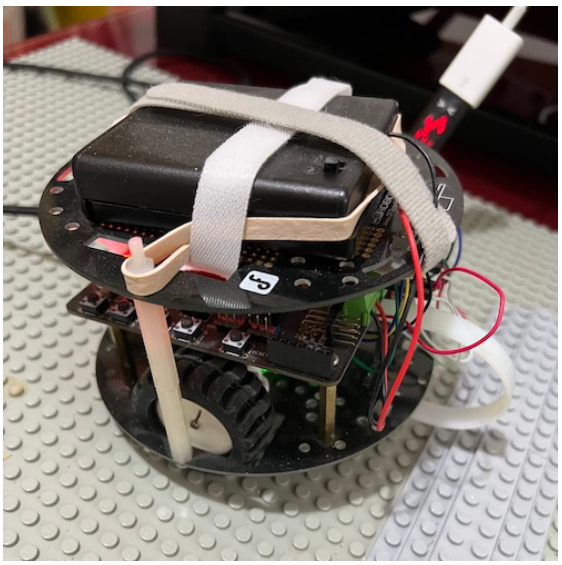

Figure 5: MATLAB controlled mobile robot

Figure 4 shows the full neural PID controller. The role of the D-circuit is to respond to fast transient changes in the error signal. For this reason, the time base, τ, of the D-circuit is set to respond more rapidly than either the P-circuit or I-circuit.

8. Robot Neural PID Control

To test the neural PID controller in a realistic environment, the controller itself is modelled in MATLAB, and this is used to control a small Arduino based mobile robot as seen in Figure 5. The only goal of the robot is to maintain a constant speed (0.5 revs/second) as measured by shaft-encoders on each wheel. The robot runs along a short (77cm) track, and it encounters a low obstacle halfway along this track that it must surmount.

The time base of the D-circuit is set to half that of the other circuits to enable rapid response to change. Each wheel has its own independent controller, though for simplicity only the values for the left wheel controller are plotted in Figure 6. Aside from the speed (revs/sec – green) the graph shows the activation levels of the input layer (error – blue) and the three PID circuits in the middle layer.

The robot begins from a standing start, causing an initial spike in the activation of the D-circuit (purple). This contributes to the initial acceleration of the robot. The P-circuit (red) isn't far behind, and after an initial peak the proportional contribution falls gradually as the I-circuit (orange) picks up the slack to find the level of activation needed to maintain speed on the flat. Between 4-5 seconds into the experiment the robot reaches the obstacle – a small Lego hump – and begins to slow down (green). The error (blue) climbs and again the D-circuit kicks in to compensate, with the P-

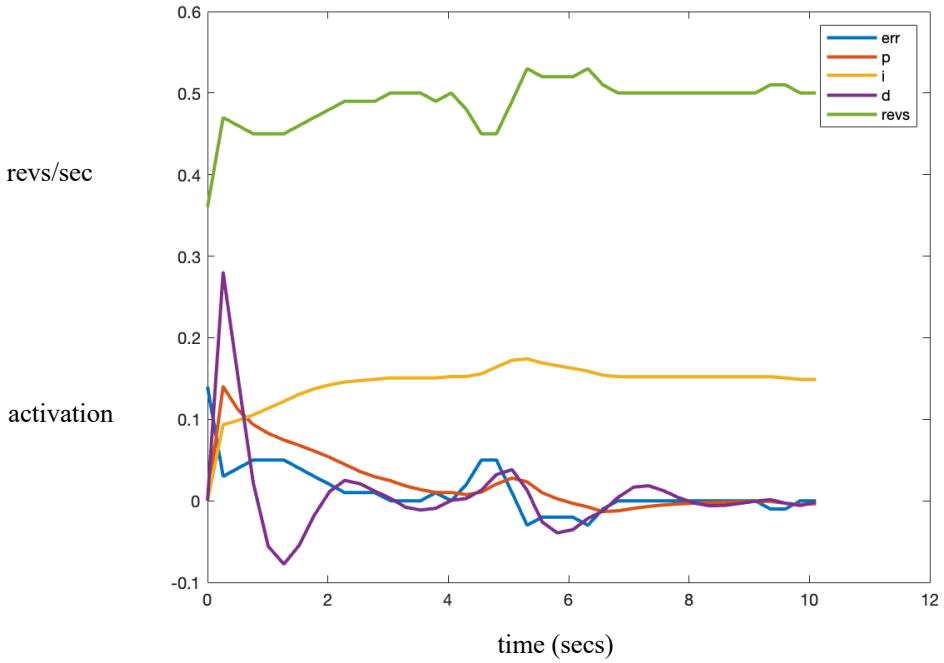

Figure 6: PID activation levels for 'obstacle' course.

circuit following closely behind. We can see the D-circuit eventually converging to a value that maintains an even speed as the error and P-circuit output falls to 0.

Tuning a PID controller in a real environment is challenging; the biggest challenge being to avoid oscillation. A simple trial and error approach was used, adding the P, I, D circuits in that order, starting by gradually incrementing the proportional weighting, K_p, until it no longer oscillates. Then incrementing the I-circuit weighting, K_i for $K_p < K_i$, until oscillations are again eliminated and it converges on the target speed. Finally, oscillations added by the D-circuit are eliminated by reducing its weighting, K_d for $K_d < K_p$. The system is found to work well where $K_d < K_p < K_i$.

9. Conclusion

This paper has presented a Continuous-Time Recurrent Neural Network (CTRNN) implementation of a PID controller that is defined in a way that doesn't rely on discrete time intervals. Experimental validation using a mobile robot confirms the practical effectiveness of the approach, with successful speed control over varied terrain.

References

Bennet, S. (1984). Nicolas Minorsky and the Automatic Steering of Ships. *Control Systems Magazine*, *November*, 10–15.

Forcada, M. L. (2002). *Discrete-time recurrent neural networks*. https://www.dlsi.ua.es/~mlf/nnafmc/pbook/node21.html

Mayr, O. (1970). *The Origins of Feedback Control*. M.I.T. Press.

Mihailoff, G. A., & Haines, D. E. (2015, June 13). Motor System I: Peripheral Sensory, Brainstem, and Spinal Influence on Anterior Horn Neurons. *Clinical Gate*. https://clinicalgate.com/motor-system-i-peripheral-sensory-brainstem-and-spinal-influence-on-anterior-horn-neurons/

Minorsky., N. (1922). DIRECTIONAL STABILITY OF AUTOMATICALLY STEERED BODIES. *Journal of the American Society for Naval Engineers*, *34*(2), 280–309. https://doi.org/10.1111/j.1559-3584.1922.tb04958.x

Powers, W. T. (1981). *Behavior, the Control of Perception*. Aldine.

Purves, D. (2008). *Neuroscience*. Sinauer.

Shu, H., & Pi, Y. (2000). PID neural networks for time-delay systems. *Computers & Chemical Engineering*, *24*(2–7), 859–862. https://doi.org/10.1016/S0098-1354(00)00340-9

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine* (2nd edn). MIT Press.

Relational Norms for Human-AI Interaction: Some Preliminary Empirical Findings

Madeline G. Reinecke, Andreas Kappes, Ilina Singh, Julian Savulescu, & Brian D. Earp

Keywords: relationships, norms, cognitive science

Abstract

Artificial intelligence (AI) systems are increasingly filling social roles traditionally filled by humans, raising questions about whether people apply the same relational norms to human-AI interactions as they do to human-human ones. To investigate this, we surveyed a gender-representative sample of U.S. adults ($N = 432$) on their expectations regarding four categories of relational norms—care, hierarchy, transaction, and mating—across seven types of relationships (i.e., fellow workers, supervisor-assistant, teacher-student, mental health provider-patient, customer-seller, long-term romantic partners, close friends). Depending on condition, we described these relationships as consisting of either of two humans or one human and a superintelligent AI. Participants then rated the extent to which each party in the relationship should—or should not—behave in a manner consistent with care, hierarchy, transaction, and mating.

Our results revealed a complex picture: While norms related to care and mating were strongly endorsed in certain kinds of human-human relationships (e.g., between romantic partners or close friends), people often resisted applying these same norms to analogous human-AI interactions. Hierarchical norms were likewise applied differently to human-AI contexts, with participants emphasizing that humans should maintain equivalent or greater authority than AI in nearly all relationship types (e.g., such that a human student should have equivalent authority to an AI teacher). These and other findings, described below, suggest that relational norms developed for human-human interactions may not transfer to human-AI relationships, highlighting the need for behavioral science to inform the design of socially acceptable AI systems.

Background: The integration of AI into the human social world—via chatbots like Replika or Xiaoice—has transformed how individuals interact with machines. These AI systems can mimic roles traditionally reserved for humans, such as personal assistants, therapists, or even romantic partners. However, the psychological and normative dimensions of these relationships remain underexplored. Do people expect to treat AI "partners" the same way they treat human partners? Can norms develop for human relationships adequately capture human-AI social dynamics?

Prior research in psychology identifies four key relational functions—care, hierarchy, transaction, and mating—as central to coordinating human-human interactions (Earp, 2021; Earp et al., 2021). These norms emerged across human societies to resolve recurrent coordination challenges faced by our species (Curry et al., 2019); but they may not seamlessly extend to interactions with AI, which lack the same biological and psychological underpinnings. Consider the following “translation” problems, highlighting key differences between humans and AI that might impact the application of each relational norm:

1. **Care:** Norms of care involve non-contingent support for another’s welfare needs, as seen in parent-child or close friendship dynamics. But AI systems do not have welfare needs on most accounts. In human-AI contexts, might participants see care as unidirectional, with the expectation that AIs should behave in a caring manner toward humans but not vice versa?
2. **Hierarchy:** Human relationships often involve hierarchies, as in manager-employee or teacher-student interactions. Yet there is reason to think humans may be resistant to ceding any type of authority to an AI (e.g., due to fear, Cugurullo &

Acheampong, 2024), even if the latter is in a traditionally hierarchical role, such as a supervisor.

3. **Transaction:** Transactional norms, based on reciprocity and exchange, are central to peer relationships. Human-AI relationships, however, are often mediated by third parties, such as AI developers or companies. This transactional framing may influence how users perceive even non-transactional roles, such as AI "friends" or “romantic partners.”
4. **Mating:** Romantic and sexual relationships are underpinned by norms of commitment, affection, and, in some cases, reproduction. While such norms may seem natural and appropriate within human-human relationships, might they be viewed as unnatural or inauthentic in human-AI romantic interactions?

In this work, we empirically investigate whether participants apply relational norms from human-human relationships to human-AI relationships, providing a foundation for future work on human-AI relational norms in both psychology and ethics.

Methodology: Participants were asked to imagine a world 10 years into the future, in “a world that… includes artificially intelligent (AI) agents.” We prompted them to evaluate relational norms across seven kinds of relationships, from which we could compare human-human to human-AI dynamics. Additional measures assessed participants’ fears about AI assuming social roles, their perceptions of the "realness" of AI relationships, and their attributions of sentience and autonomy to AI systems.

Key Findings and Implications: On the whole, human-AI relational norms varied systematically from human-human relational norms. There was no single category of norms (e.g., hierarchy) that was entirely aligned or misaligned between human-human and human-AI cases. Even in the domain of mating, where norms varied most widely between human-human and human-AI relationships, we observed similarity in some instances (e.g., in denying romantic behavior as appropriate for the teacher-student relationship; K-S = .08, Bonferroni-corrected p = 1.00). Moreover, no single human-AI relationship entirely aligned (or misaligned) with its human-human equivalent. Even for romantic partnerships, where participants distinguished most strongly between human-human and human-AI pairs (e.g., in terms of how appropriate romantic behavior would be), there was still marked similarity in some areas (e.g., transaction, K-S = 0.15, Bonferroni-corrected p = 1.00). Overall, however, we observed a trend towards resisting care and mating in human-AI relationships, which rely on mutual emotional investment. This suggests a potential barrier to the social acceptance of AI in intimate or emotionally significant roles. Next, we also observed a tendency for participants to endorse norms that would maintain human authority over AI — even in collaborative or peer-like relationships, and even in relationships in which the human occupied the traditionally subordinate role (e.g., a human student and an AI teacher). Again, this suggests that human-human norms may not successfully transfer to human-AI relationships.

Ethical Design Considerations: As AI increasingly occupies social roles, understanding how relational norms translate—or fail to translate—can inform the development of AI systems that align with human values and expectations.

Conclusion: Our findings underscore the complexities of applying human relational norms to human-AI interaction. As AI continues to evolve, it is imperative to consider these dynamics to foster ethical, socially appropriate integration into human society.

References

Earp, B. D. (2021). *Relational Morality* (Doctoral dissertation, Yale University).

Earp, B. D., McLoughlin, K. L., Monrad, J. T., Clark, M. S., & Crockett, M. J. (2021). How social relationships shape moral wrongness judgments. *Nature Communications*, *12*(1), 5776.

Cugurullo, F., & Acheampong, R. A. (2024). Fear of AI: an inquiry into the adoption of autonomous cars in spite of fear, and a theoretical framework for the study of artificial intelligence technology acceptance. *AI & Society*, *39*(4), 1569-1584.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47-69.

The Subject Project: Human Visibility, Vulnerability, and Diversity amidst Proxy Discrimination in the Data Age

James Garrison
University of Massachusetts Lowell, Lowell, USA

Abstract

Using a combination of approaches from critical theory, data science, and data ethics, my upcoming monograph *The Subject Project: Human Visibility, Vulnerability, and Diversity in the Data Age* investigates 1) how human subjects view themselves and are conscious of being viewed by data systems and 2) how to mitigate the impact that this has, particularly on vulnerable populations. *The Subject Project* maintains that specific conditions of visibility and invisibility make each of us (but some much more so than others) vulnerable to being compelled to see what we hold to be special about ourselves projected with unsettling accuracy into the future by impersonal algorithms which have major social and political consequences.

The Subject Project thus builds on perspectives from critical theory like those of Michel Foucault regarding the emergence in the industrial age of self-monitoring, self-policing, self-disciplining, self-conscious subjects being formed by the gaze of others in society's "virtual panopticon." Now, with the advent of the data age, the human subject has also become a project, i.e., a projection. However, much like subjection leads to disparate impact on particularly visible and vulnerable subjects, diverse populations are projected as data in ways that lead to real harm. How so?

Consider how machines learn correlations in multi-dimensional datasets and flatten them into simpler human readable projections in domains like finance, medicine, and law enforcement. Even if a noxious category (e.g., race and/or sexual orientation) is removed from databases, there are often proxies in public sources sufficient to make predictions of an extensionally defined population who just happen to be ... all of the same minority category.

For example, simple facial portraits already provide mathematical data that can be used to predict sexuality, stoking fears where non-heterosexual activity is punished. Moreover, machine learning provides useful output for decision making commensurate to the quality of its input data. Hence, with a more robust data set, correlations can be established that effectively de-anonymize supposedly anonymous persons. You might sign user agreements with boilerplate language regarding anonymous data usage, but do you really feel confident that so many other people exist in your postal code who share your map searches/transportation habits, your fitness app scores, your search history for symptoms, your hyper specific romantic/sexual tastes, etc. that you actually enjoy meaningful anonymity/privacy and will not be vulnerable to punishment for being visible while living your life. Such data proxies can, and already are, being used, to guide actions and policies harmful to minority groups and further entrenching disadvantage. This is like updating America's "Jim Crow" laws with digital "Jim Code."

Accordingly, I call for policies prohibiting digital segregation using proxy categories. Just because artificial intelligence can evade established laws by projecting and predicting typically protected demographic and financial information without making use of forbidden data, hard-won anti-discrimination protections *must not* be diminished *ever*.

The Future of Technology and How Mathematics is Implemented: initial workshop outcomes

Marie Oldfield (School of Mathematics, LSE, London, UK)

1 INTRODUCTION

A number of ethical AI frameworks [1] have been published designed to guide developers involved in producing AI systems in order to help to mitigate the risks and harms which can occur. Only systems developed along such ethical guidelines can be considered "Trustworthy AI".

It is clear that AI systems which infringe on equality and human rights, and demonstrate significant bias continue to be deployed despite the prevalence of these guidelines [2]. Court cases have now started around AI impact [3], its legislation and regulation[4] [5]. However, two problems exist here:

- Firstly, the notion that one can hold practitioners to account without having them been properly training them is unethical.
- Secondly, if either the law or policy has not been translated so that it can be implementational means the practitioner, in question, cannot adhere to that law or policy.

This gap was viewed as so severe that it was determined, after significant empirical research [6] [7], that a professional body and professional accreditation were needed by practitioners in AI. This provision would help practitioners display their knowledge, provide agreed ways forward on best practice, provide a voice for practitioners and support practitioners with their responsibilities. In essence this is the standardisation of a discipline which The Institute of Science and Technology, (IST), holds independently developed the AI Professional Accreditations for AI Practitioners (approved by partners in academia, industry and government via steering board) and Educational Quality Marks for higher education and commercial training providers.

The professional accreditation was developed in conjunction with industry, academia and government partners. The accreditation starts at entry level (RTechAI) of an individual, moving to Mid-Career Level (RPAI), and then to Advanced Practitioner (APAI). The IST educational quality marks ensure that higher education courses and commercial courses meet the rigorous standards expected of practitioners. However, the link between academia and technology is still not functioning as it should [8] and this is leading to debates, such as those at this research school, about who is accountable, who is responsible and who should bear the burden of the law. Clearly no one wants to take on this burden but frankly this burden belongs to the entire chain of development.

The question exists why sectors of society are reluctant to take on this burden and why. Academia, as a sector, is reluctant to take on the responsibility on the grounds that the theory those in this sector generate could be used in the wider world. There is also reluctance to caveat this work for safety. This illustrates a gap in best practice. Theory becomes practice and it is not addressed where either accountability or responsibility of the transfer of theory transfer to industry from academia. Bluntly industry cannot be expected to shoulder this burden alone.

2 THE ROLE OF EDUCATION

Academia, as a sector, has a responsibility to educate future practitioners both as to the ethics and responsible and robust implementation of AI. The UK government has provided benchmark statements, produced by the Quality Assurance Agency for Higher Education (QAA) which formalise the teaching and assessment of taught courses in secondary and higher education [9]. Ethics is currently not included within the subject benchmark statements governing secondary and tertiary education [10]. The QAA have updated the benchmark statements recently but it did not include either any further ethics or best practice within the benchmark statements despite this being requested.

At the same time, it is clear AI teaching and assessment is neither formalised orn overseen. With the rise of legislation and regulation practitioners may be held to account but at the same time lack a formalised and benchmarked education. This education does not necessarily get "picked up" by workplace training. Unfortunately, many employers in the corporate sector appear not to place great emphasis on workplace training. The Institute of Science and Technology has looked to fill this gap by accrediting educational and commercial courses across the UK in AI with a quality mark. This quality mark enables allows students to enter straight into the student AI designation at the Institute of Science and Technology. The quality mark assures the quality and content of the course. With little oversight of AI training and teaching in the UK students currently cannot make an informed choice on content or quality of course.

3 THE PROPOSALS

- Proposal 1: We need formal benchmarked AI and Ethics teaching and assessment in Schools and Universities
- Proposal 2: The IST AI Professional Accreditation is needed as an industry standard

4 STAKEHOLDER WORKSHOP

To assess the acceptability and feasibility of the proposals, we conducted a research school funded by the London Mathematical Society with key stakeholders. We aimed to discover what gaps students believe exist within the transition from theory to practise. Academic staff and representatives from industry attended and were part of the discussions. Technical and nontechnical talks were held with the specific goal to stimulate discussion. The students then worked with lecturers and industry professionals to

examine and explore the issues they face when developing theory which then goes on to be implemented in the real world.

Participants, throughout the two-day research school, posted post it notes with their thoughts, questions and statements:

- Question 1 – What gaps exist when implementing theory into practise?
- Question 2 – What further information do you need to overcome obstacles in implementing your work in the real world?
- Question 3 – What communication issues exist for you, as an individual, in order to effectively translate your work into the real world in robust and safe way?

It was clear that students were very engaged in the industry – academia connection and they recognised that once faced with future job roles that they ought to be aware of the potential regulation and legislation. They recognised from their perspective that as a student went into either industry or academia and designed an algorithm which was unethical, they are faced the following issues:

- How do I know if this algorithm is unethical or not?
- Why am I not being taught ethics of AI?
- What happens to me if there is an adverse effect as a consequence of my work using AI algorithms?
- How can I learn to model ethically?

This was reflected in longer form feedback in Table 1.

Participant	Comment
1.	What should be taught? Data Source Data Use Outcomes What is AI
2.	How does one implement ethical AI and not just talk about implementing it
3.	Thinking of what data doesn't get fed into the AI such as oral data of native populations
4.	Regulation of reinforcement learning via critical evaluation of policy Independent consumer protection by experts
5.	In which way (where) to educate people about AI; is it at university, school or whilst in the workplace?
6.	Science communication: Explain simple algorithms/ideas ie make people invested Give people a feeling of advocacy and initiative Make life difficult for unethical behaviour
7.	Having a development pipeline for AI models that add emphasis to data collection and planning of model topology is essential to ensure audit trails are clear
8.	Concrete steps are needed for applying trustworthy ethical AU assessments
9.	Have a regulatory body with robust policies and regulations on AI in education
10	Invest in diversifying AI. Community: this has a direct impact because a more diverse community will be better equipt to spot and review biases
11.	Modelling data and creating AI models How to check the model is good enough to be shared
12.	What sort of ethics can we teach to our students
13.	Why students are struggling in producing a good ML or AI model
14.	Advocates within the research fields as multiplicators Student initiatives ad involvement in SIAM IST Student chapters

Table 1. Longer Form Feedback

5 CONCLUSION

Many of the comments obtained in the workshop echoed the preceding papers by Holstein et al. [11], Oldfield et al. [12] [13] and Oldfield [14]. It is clear that since 2018 there has been little improvement for practitioners and little improvement in practice since the 1990s. This is a problematic situation given that problems in AI development, as detailed above, have been consistently raised and ignored by the QAA and the UK Government.

The generation of non-granular legislation and regulation has further compounded problems as a consequence of the move towards accountability in the UK without either having correctly and formally trained our practitioners. Without a substantial move on the part of the QAA, neither teaching nor assessment cannot be aligned to the global market and workplace. Both assessment and teaching has to be aligned to the global market and workplace. Therefore, the IST accreditation is currently the only pathway to practitioner training and best practice. Lawyers are currently struggling to prosecute AI models which are unethically used and applied because they are not being specialists in this area. Once again, the sole means by which this can be done is to undertake the reverse engineering of the IST best practise pipeline to isolate (identify) gaps.

It is worrisome that little movement has taken place within the last 30 years to move best practice and education forwards but the positive aspects of this is that a professional body has been able to move the conversation forward and found an entire ecosystem dedicated to AI practitioners and form a community and standards for a discipline.

REFERENCES

[1] House of Lords AI in the UK: No Room for Complacency House of Lords Liaison Committee.

[2] BBC, Facial Recognition Use by South Wales Police Ruled Unlawful.

[3] Generative AI – IP Cases and Policy Tracker | Mishcon de Reya

[4] AI Regulation.

[5] EU AI Act.

[6] Holstein et al., Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?

[7] Oldfield, Technical Challenges and Perception.

[8] Oldfield and McMonies, Call for Written Evidence Risk Assessment and Risk Planning. Lords Select Committee, London.

[9] Oldfield, Towards Pedagogy Supporting Ethics in Modelling.

[10] Oldfield.

[11] Holstein et al., Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?

[12] Oldfield and Haig, Analytical Modelling and UK Government Policy.

[13] Oldfield and McMonies, 'Call for Written Evidence Risk Assessment and Risk Planning. Lords Select Committee, London.

[14] Towards Pedagogy Supporting Ethics in Modelling, by Marie Oldfield.

Integrating Multi-Omics Data with Deep Learning for Predictive Modelling of Autoimmune and Neurodegenerative Diseases

Priya Gopal [1] , **Solaimurugan Vellaipandiyan** [2] , **Nazi Mohammed Navi** [3]

Abstract. In precision medicine, predicting the onset and progression of autoimmune and neurodegenerative diseases poses a multifaceted challenge due to their intricate molecular complexities. Traditional diagnostic methods, while foundational, often fall short in capturing the full scope of these dynamic and heterogeneous pathologies. This study introduces an advanced framework that integrates multi-omics data, including genomics, transcriptomics, proteomics, and metabolomics—to enhance predictive modeling for such conditions. Leveraging the power of sophisticated deep learning methodology allows the framework to establish patterns and interdependencies at a scale greater than trivial for several layers of biological strata. The Feedforward model within this framework demonstrates exceptional predictive performance, achieving an accuracy of 97.34%, an F1 score of 97.20%, and a balanced accuracy of 97.45%. The incorporation of explainable AI (XAI) techniques ensures model interpretability, facilitating the identification of novel biomarkers and revealing previously obscured pathways of disease progression. This integrative approach not only advances the precision of prognostic models but also catalyzes the development of personalized diagnostic and therapeutic strategies, marking a significant milestone in the evolution of precision medicine.

Keywords: Multi-Omics Integration, Deep Learning, Autoimmune Diseases, Neurodegenerative Disorders, Explainable AI, Predictive Modeling

1 Introduction

The advent of multi-omics technologies encompassing genomics, transcriptomics, proteomics, and metabolomics has profoundly enhanced our comprehension of complex diseases by offering an integrated, systems-level perspective of molecular mechanisms [1, 2]. Among these, transcriptomics, which examine gene expression profiles, has been pivotal in elucidating how genes are regulated under specific conditions. This capability is particularly valuable in the study of autoimmune diseases, such as rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE), as well as neurodegenerative disorders, including Alzheimer's disease (AD) and Parkinson's disease (PD) [3]. These diseases are characterized by intricate, multi-layered interactions between genetic, transcriptional, and metabolic factors, presenting significant challenges for traditional diagnostic and predictive methods [4].

While substantial progress has been made in leveraging single-omics data to explore disease mechanisms, the integration of multi-omics data remains a formidable challenge due to its inherent heterogeneity and high dimensionality [5]. Traditional statistical approaches often fall short in capturing the complex, non-linear relationships between different layers of omics data, limiting their applicability in clinical diagnostics and personalized medicine [6]. Recent advances in artificial intelligence, particularly deep learning (DL) [7], have demonstrated great promise in addressing these challenges. DL architecture excels in modeling complex patterns and dependencies, making them well suited for multi-omics integration.

This study introduces a novel deep learning framework designed to integrate multi-omics data with a specific focus on transcriptomics, aiming to improve the prediction and understanding of autoimmune and neurodegenerative diseases. The proposed framework employs state-of-the-art architectures, including bidirectional long short-term memory (BiLSTM), gated recurrent units (GRU), and feed-forward neural networks, to capture the temporal and spatial dependencies inherent in omics data. To enhance clinical applicability, the framework incorporates explainable AI (XAI) techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), enabling the identification of key biomarkers and pathways that drive disease mechanisms. [32]

By focusing on autoimmune and neurodegenerative diseases, this study bridges the gap between computational innovation and clinical relevance. The integration of multi-omics data through the proposed framework not only enhances predictive accuracy but also provides actionable biological insights, paving the way for early diagnosis and personalized treatment strategies. This dual advantage superior predictive accuracy and actionable biological insights positions the framework as a potential benchmark in the field, offering a transformative approach to disease modeling and precision medicine [9].

2 Literature Review

Multi-omics data integration, which embraces genomics, transcriptomics, proteomics, and metabolomics, has significantly advanced our understanding of human pathophysiology by providing a panoramic view of molecular networks [10,11] . Among the available tools, transcriptomics, especially gene expression profiling has been instrumental in revealing disease mechanisms by portraying the regulatory dynamics of the activation and suppression of genes as a response to genetic and environmental stimuli. The significance of this

[1] University of West England, Bristol, United Kingdom, email: priya.gopal@uwe.ac.uk
[2] Centre for Development of Advanced Computing Chennai, Chennai, India, email: solaimurugan.v@gmail.com
[3] Faculty of Computer Science and Information Technology, University of Tun Hussein Onn Malaysia, email: nazri@uthm.edu.my

is most needed in multifactorial diseases like autoimmune diseases, rheumatoid arthritis, and systemic lupus erythematosus; neurodegenerative disorders such as Alzheimer's disease; and Parkinson's disease [34], where genetic, epigenetic, transcriptomic, and metabolic alterations are intertwined, creating intricacy in both diagnosis and prognostication.

Conventional diagnostic and predictive methodologies are inadequate in capturing the complex and heterogeneous nature of these diseases, particularly within autoimmune and neurodegenerative domains. Gene expression data serves as dynamic reservoirs of biolog- ical information, yet the integration of gene expression data with other omic modalities such as proteomics and metabolomics remains a daunting challenge. These omic datasets are intrinsically high- dimensional, noisy, and heterogeneous, so quite elaborate computa- tional frameworks are required in order to capture these subtle inter- actions between layers. According to [12], a considerable interest in the diagnostic applications of multi-omics data lies with the integra- tion of XAI approaches. Indeed, SHAP (Shapley Additive Explana- tions) and LIME (Local Interpretable Model-Agnostic Explanations) are key enabling techniques toward improved model interpretability, the key issue for model transparency that significantly impacts their potential for adoption in clinics. In a systematic fashion, [8] mapped many XAI methodologies applied to multi-omics data by focusing on a fast evolving need to develop transparent and interpretable AI mod- els applicable in biomedicine. The potential of multi-omics integration for precision medicine has been the subject of extensive investigation. [13] demonstrated how multi-omics data can be synergistically employed in Parkinson's disease to refine therapeutic strategies, while [31] explored the integration of single-cell and multi-omics datasets to uncover novel drug targets in autoimmune disorders, suggesting that deeper mechanistic insights can inform the development of targeted treatments. The combined analysis of transcriptomics, proteomics, and metabolomics, as discussed by [15], holds transformative potential for personalized therapeutic strategies, especially in neurodegenerative diseases like Alzheimer's, where precision medicine approaches are increasingly being advocated. Recent strides in computational techniques have leveraged deep learning models to manage and integrate multi-omics data effectively. AUthor [16, 17] looked into the challenges and opportunities arising from multi-omics integration for personalized healthcare, emphasizing how deep learning architectures represent complex interrelations between heterogeneous omic layers. The authors in [18] applied deep learning models [19], such as BiLSTM (Bidirectional Long Short-Term Memory) and BiGRU (Bidirectional Gated Recurrent Units), to predict hematologic malignancies like leukemia, showing the potential of deep learning techniques in facilitating multi-omics integration. These architectures show great potential in handling sequential and multidimensional datasets a feature typical of biological systems and therefore are well applied in disease prediction, just as [20, 21] did using multi-task learning frameworks for predicting autoimmune diseases. Researchers, such as [14] and [22], have recognized the transformative potential of deep learning for multi-omics-based disease prediction. The GREMI framework, proposed by [33], highlighted the importance of XAI in optimizing multi-omics integration for improving the accuracy of disease prediction while maintaining model transparency. Likewise, [23] proposed TMO-Net, a pre-trained, explainable multi-omics model using multi-task learning for oncology applications, further solidifying the need for interpretability in clinical AI models [24]. In a comprehensive survey, [25,26] reviewed deep learning models for multi-omics analysis, highlighting their promise in disease prediction and classification but also bringing forth the challenges in achieving high levels of interpretability and scalability in clinical contexts. These advances show that deep learning methodologies are poised to extract meaningful insights from multi-omics datasets, but overcoming the interpretability barrier remains a critical challenge for widespread clinical adoption. Despite these advances, challenges remain in the integration of multi-omics data for clinical decision-making. [27, 28] benchmarked several multi-omics integration methods in precision medicine and called for more robust validation and cross-validation protocols in model development. Moreover, [29] highlighted that the application of XAI techniques to multi-omics data is still at an early stage, requiring further exploration to establish the reliability and credibility of these models in clinical settings. Despite their impressive predictive power, deep learning models are still difficult to interpret, and this is a significant obstacle to their acceptance in healthcare settings. This study shows an adaptive deep learning framework designed to integrate multi-omics data, with a specific focus on gene expression profiling for disease prediction. The model employs advanced architectures such as BiLSTM, BiGRU, recurrent neural networks (RNNs), and feed-forward neural networks (FFNNs) to capture both temporal and spatial dependencies in the data. To enhance the model's clinical applicability, explainable AI techniques such as SHAP and LIME are incorporated to ensure accurate and interpretable predictions. By improving predictive accuracy and providing deeper insights into disease mechanisms, this approach sets a benchmark for disease modeling, facilitating early diagnosis and personalized therapeutic interventions.

3 Methodology

3.1 Raw Data

The study utilizes multi-omics datasets from well-known public repositories, such as Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) [30] which include a wide variety of genomic sequences, gene expression profiles, proteomic data, and metabolomic information. By combining high-dimensional, heterogeneous datasets, the model enhances the accuracy of disease prediction and aids in the identification of precise biomarkers, thereby deepening the understanding of autoimmune and neurodegenerative diseases.

3.2 Data Extraction and Cleaning

Gene Expression Data: We extract gene expression data for each disease from the unstructured raw GEO Series Matrix files. The expression values are logged-transformed to ensure normal distribution of gene expression levels, which is critical for model performance.

Annotation Integration: Gene annotation information, such as gene symbols, is integrated with the expression data based on probe identifiers, which are unique codes (e.g., GSM1116933, GSM1116935, GSM1116947) used to identify specific genes in the dataset. This integration ensures accurate mapping of gene expression data to biologically meaningful identifiers, such as gene symbols or Ensembl IDs. By merging annotation with expression data, a comprehensive dataset is created, where gene symbols are directly linked to their corresponding expression values across biological samples, enabling more meaningful and interpretable analyses of gene activity, molecular pathways, and biological processes. The top 10 gene symbols for each disease dataset are visualized in Figures [1], [2], [3], and [4], representing the proportional expression levels of key genes in Rheumatoid Arthritis, Alzheimer's Disease, Parkinson's

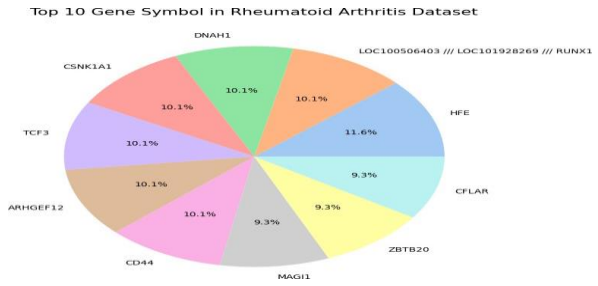

Figure 1: Top 10 Gene Symbols in Rheumatoid Arthritis Dataset

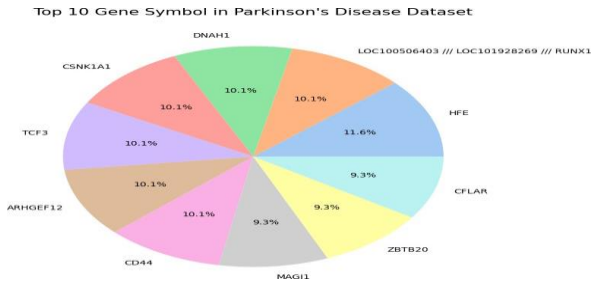

Figure 2: Top 10 Gene Symbols in Parkinson's Disease Dataset

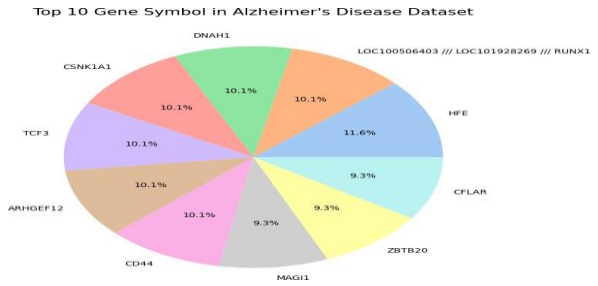

Figure 3: Top 10 Gene Symbols in Alzheimer's Disease Dataset

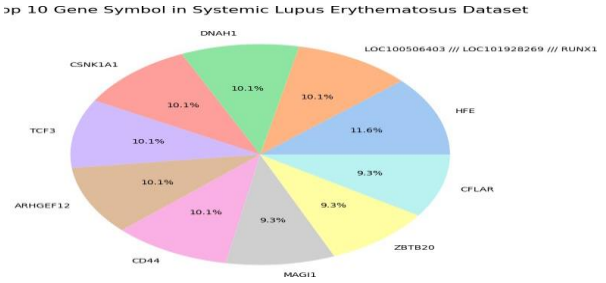

Figure 4: Top 10 Gene Symbols in Systemic Lupus Erythematosus Disease Dataset

Disease, and Systemic Lupus Erythematosus (SLE). These figures provide a comprehensive view of the molecular underpinnings of each disease, highlighting genes that play significant roles in pathophysiology.

Rheumatoid Arthritis (RA) (Fig. [1]): In the RA dataset, HFE, associated with hemochromatosis, exhibits the highest expression level, suggesting its critical role in iron metabolism dysregulation within the inflammatory process of RA. ARHGEF12, involved in Rho GTPase signaling, may influence cellular dynamics and immune response regulation. CD44, a cell adhesion molecule, plays a pivotal role in immune cell migration and tissue inflammation, which is significant in RA pathogenesis. MALAT1, a non-coding RNA, is known to regulate gene expression and could impact synovial inflammation in RA.

Alzheimer's Disease (AD) (Fig. [3]): In Alzheimer's Disease, HFE is again prominently expressed, indicating a potential link between iron overload and neurodegenerative processes. The involvement of ARHGEF12 in neuronal signaling suggests a role in synaptic dysfunction. CD44, which influences amyloid plaque formation, is significant in AD, potentially affecting neuroinflammation. TCF3, a transcription factor, could be involved in the regulation of genes that modulate neuronal survival and plasticity, contributing to AD progression.

Parkinson's Disease (PD) (Fig. [2]): The PD dataset reveals HFE's role in iron metabolism, which is critical in dopaminergic neuronal health. ARHGEF12, through its effects on cellular signaling, may modulate neuroinflammatory pathways. CD44, again critical in immune cell interaction, could influence neuroinflammatory responses in PD. DNAH1, which encodes a dynein protein, is involved in cellular transport and may contribute to neurodegeneration by impairing axonal transport in PD.

Systemic Lupus Erythematosus (SLE) (Fig.[4]): For SLE, HFE's involvement in iron regulation and immune modulation underscores its significance in autoimmune diseases. ARHGEF12 and its influence on immune signaling pathways may contribute to the aberrant immune responses seen in SLE. CFLAR, an apoptosis regulator, could play a role in controlling T-cell death and maintaining immune system balance, a critical aspect of SLE pathology. The presence of MALAT1 suggests a potential role in modulating immune cell function and the inflammatory response in SLE. In each of these diseases, the expression of these genes is tightly linked to the immune system dysregulation, cellular signaling, and neuronal degeneration, emphasizing their potential as therapeutic targets for classification.

3.3 Data Preprocessing

Missing Values Handling: Missing values in gene expression data are imputed using appropriate statistical methods such as KNN imputation or mean/mode imputation depending on the nature of the data. **Filtering:** Only genes with non-missing gene symbols and those that occur in at least 5 samples are retained, ensuring that low-expressed genes, which could introduce noise into the models, are excluded from further analysis. **Data Normalization and Scaling:** Prior to training, the data undergoes normalization and scaling. Feature scaling is performed using methods such as StandardScaler or MinMaxScaler to standardize the data, ensuring that all features (genes) are within a comparable range. This is crucial for the performance of neural networks, which are sensitive to feature scaling. **Class Balancing:** Given the imbalanced nature of disease classification, where certain diseases have fewer samples, class balancing techniques like Synthetic Minority Over-sampling Technique

(SMOTE) are applied to ensure that the model receives an equal distribution of samples from each class. **Dimensionality Reduction:** Principal Component Analysis (PCA) or other dimensionality reduction techniques are applied to reduce the high-dimensional gene expression data into a manageable set of features, while retaining the variance that is most informative for disease classification. **Correlation Matrix:**The correlation with gene symbols in various datasets is illustrated in Figures [5] for Rheumatoid Arthritis, [6] for Parkinson's Disease, [7] for Alzheimer's Disease, and [8] for Systemic Lupus Erythematosus. These correlation matrices highlight the significant relationships between gene expressions in each disease dataset. Understanding these correlations is critical in the context of multi-class classification, where genes that exhibit strong correlations can provide crucial features for differentiating between disease classes, improving model accuracy, and enhancing the prediction of disease-specific patterns. In the RA dataset fig. [5], GSM1116954 shows the strongest correlation, followed by GSM1116933 and GSM1116956, which also exhibit significant positive correlations. These genes may play a crucial role in the immune response and inflammation processes central to the pathogenesis of RA. In a multi-class classification context, these genes can help distinguish RA from other diseases by capturing unique patterns of gene expression, facilitating better disease classification and early detection. For the PD dataset fig. [6], GSM184364 shows the highest correlation, with GSM184370 and GSM184354 following closely. These genes are potentially involved in neurodegenerative processes and cellular signaling pathways crucial for the progression of Parkinson's Disease. Their inclusion in a multi-class classification model allows for more accurate identification of Parkinson's Disease compared to other neurological conditions, improving diagnostic precision. In the AD dataset fig. [7], GSM1303169 exhibits the highest correlation, followed by GSM1303172 and GSM1303170. These genes are likely associated with neuronal health and cognitive decline, playing important roles in Alzheimer's pathology. In multi-class classification, these genes help differentiate Alzheimer's Disease from other neurodegenerative disorders by providing essential markers of cognitive impairment and neurodegeneration. In the SLE dataset fig. [8], GSM1256978 shows the strongest correlation, followed by GSM1256974 and GSM1256973. These genes are likely involved in immune dysregulation, a key factor in SLE's inflammatory processes and immune response. Their inclusion in multi-class classification models allows for better differentiation of SLE from other autoimmune diseases, improving predictive accuracy and understanding of disease-specific gene expression patterns.

Data Splitting: The dataset is divided into training and testing subsets using an 80-20 split returning the following distributions in Table [1]:. The training data is used for model development, while the testing set is reserved for performance evaluation. Cross-validation techniques, such as K-fold cross-validation, are applied during training to ensure generalization and mitigate overfitting.

3.4 Modeling, Training, Optimization, and Evaluation

[1] Deep Learning Models: In this study, deep learning architectures are employed for the classification of gene expression data. Specifically, the following models are explored: **Feedforward Neural Networks (FNN)**: A basic architecture for capturing non-linear relationships between genes. This simple yet powerful model allows for the initial learning of complex, non-linear patterns in gene expression data. **Sequential Recurrent Neural Network (SequentialRNN)**: SequentialRNN employs stacked SimpleRNN layers to model temporal dependencies in sequential data. While it focuses on short-term patterns, its architecture effectively captures temporal dynamics in gene expression data. The model integrates dense layers for classification, leveraging its recurrent structure to process sequential patterns efficiently. **Bidirectional Long Short-Term Memory (BiLSTM)**: BiLSTM is employed to capture both forward and backward dependencies in gene expression data. Given that gene interactions might not strictly follow sequential patterns but could also exhibit long-range dependencies, BiLSTM helps uncover complex relationships by considering both past and future interactions. **Bidirectional Gated Recurrent Units (BiGRU)**: BiGRU, similar to BiLSTM, is introduced to model both forward and backward temporal dependencies while being more computationally efficient than LSTM-based architectures. BiGRU helps in capturing complex gene expression relationships with fewer parameters, making it a faster alternative without sacrificing performance. These models are implemented using the Keras framework, with the TensorFlow backend. Each model's architecture is optimized for the gene expression datasets, with varying numbers of layers and neurons based on performance evaluation metrics.

[2] Training & Optimization: The models are trained using the Adam optimizer with a learning rate scheduler to dynamically adjust the learning rate during training, ensuring efficient convergence. The batch size is set to 32, and training is conducted for a maximum of 100 epochs, with early stopping based on convergence behavior to prevent overfitting.The high-level data flow architecture for the entire process is depicted in Figure [9]. **Hyperparameter Tuning :** A systematic approach utilizing grid search or random search is employed to optimize hyperparameters such as learning rate, batch size, number of hidden layers, and the number of neurons in each layer. This process ensures that the models achieve the best performance given the complexity of gene expression data. **Regularization:** To mitigate overfitting, dropout with a rate of 50% and L2 regularization are incorporated into the models. These techniques help prevent the network from memorizing the training data, especially in deep neural networks where high model complexity could lead to poor generalization. **Class Weighting :** To address the class imbalance problem, weighted loss functions are applied, where higher weights are assigned to underrepresented classes. This penalizes misclassifications of rare classes more heavily, ensuring the model is sensitive to all disease categories. **Early Stopping:** To further prevent overfitting, early stopping is employed. The model's performance on the validation set is monitored, and training halts once the validation performance shows no improvement over a set number of epochs (patience). This prevents unnecessary computations and ensures the model does not overfit to the training data.

[3] Evaluation: After training, the models' performance is evaluated using a separate test dataset. Multiple metrics are used to assess the models' effectiveness in classifying gene expression data: **Accuracy:** Proportion of correct predictions across all classes. Used to assess overall model performance but can be misleading in imbalanced datasets. **Precision:** Ability to correctly identify true positive instances for each class. Important for minimizing false positives, especially in high-cost domains like medical diagnoses. **Recall (Sensitivity):** Ability to identify all relevant instances of

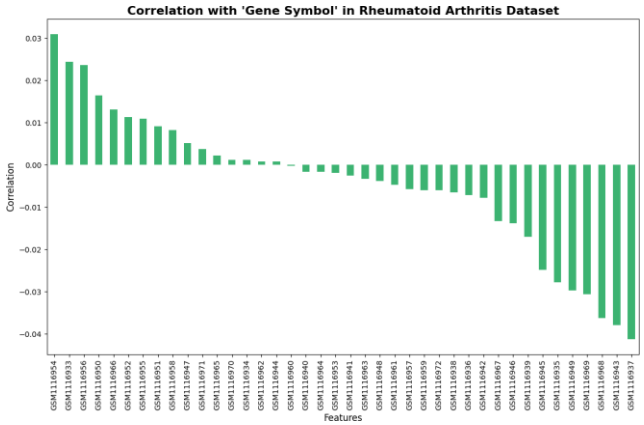

Figure 5: Rheumatoid Arthritis

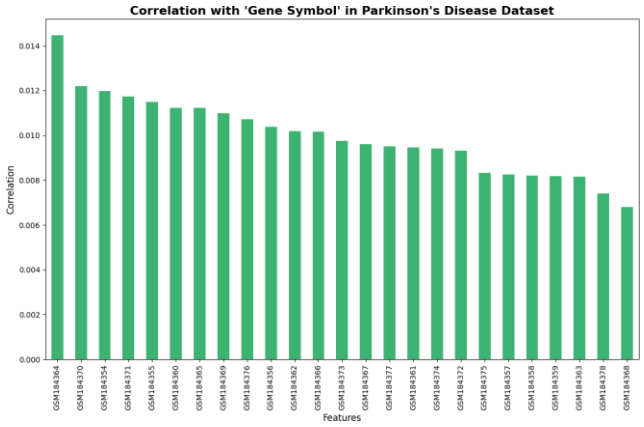

Figure 6: Parkinson's Disease

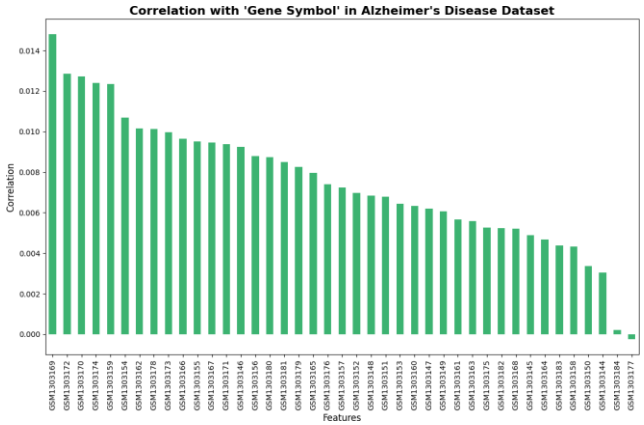

Figure 7: Alzheimer's Disease

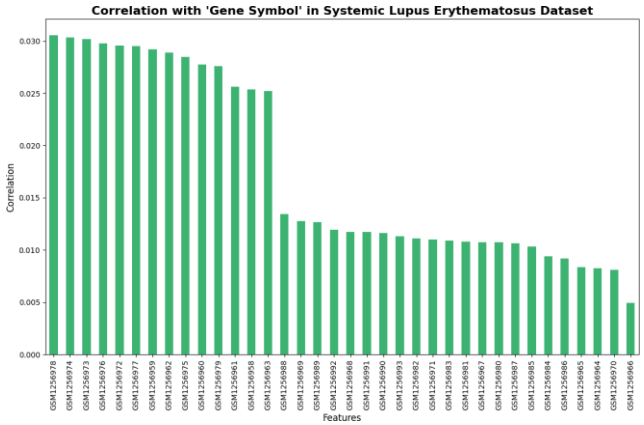

Figure 8: Systemic Lupus Erythematosus

Table 1: Data Distribution Table

Disease	Data Split	Description	Number of Records	Number of Features
Alzheimer's Disease	Total Data	Pre-processed and Over-sampled Dataset	21,707	41
	X_train	Training Feature Set	20,130	41
	y_train	Target Set for Training	20,130	-
	X_test	Testing Feature Set	1,577	41
	y_test	Target Set for Testing	1,577	-
Parkinson's Disease	Total Data	Pre-processed and Over-sampled Dataset	21,707	25
	X_train	Training Feature Set	20,130	25
	y_train	Target Set for Training	20,130	-
	X_test	Testing Feature Set	1,577	25
	y_test	Target Set for Testing	1,577	-
Rheumatoid Arthritis	Total Data	Pre-processed and Over-sampled Dataset	21,707	40
	X_train	Training Feature Set	20,130	40
	y_train	Target Set for Training	20,130	-
	X_test	Testing Feature Set	1,577	40
	y_test	Target Set for Testing	1,577	-
Systemic Lupus Erythematosus	Total Data	Pre-processed and Over-sampled Dataset	21,707	36
	X_train	Training Feature Set	20,130	36
	y_train	Target Set for Training	20,130	-
	X_test	Testing Feature Set	1,577	36
	y_test	Target Set for Testing	1,577	-

each class. Crucial for capturing true positives, especially in rare disease detection. **F1-Score:** Harmonic mean of precision and recall. Balances precision and recall, useful when both false positives and false negatives are important. **Balanced Accuracy:** Average accuracy across all classes, accounting for class imbalance. Provides a fair evaluation in multiclass problems with imbalanced data. **Kappa Score (Cohen's Kappa):** Agreement between predicted and true labels, adjusted for chance. Evaluates model reliability, especially in imbalanced or uneven class distributions.

[4] Explainable AI (XAI) Methods: To augment model transparency

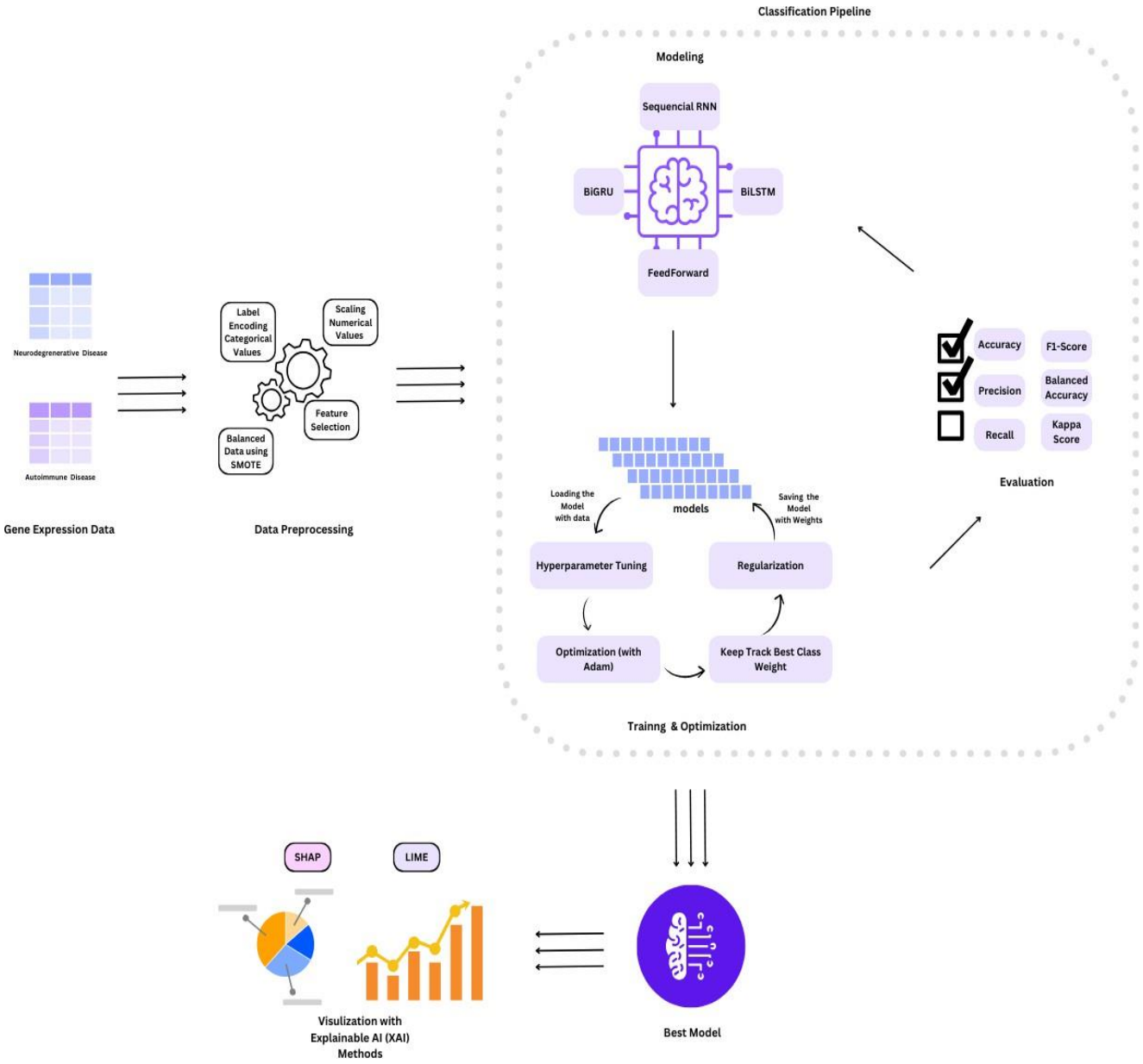

Figure 9: Data Flow Architecture for Multiclass Classification

and elucidate complex decision-making processes, XAI methodologies such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are systematically incorporated. These techniques facilitate the interpretation of intricate deep learning models and reveal the underlying pathophysiological mechanisms driving gene interactions. **LIME:** LIME operates by locally approximating high-dimensional, non-linear models with interpretable surrogate models. This enables the identification of influential features or genes for specific predictions, thereby enhancing the interpretability of model outcomes and aiding in the discovery of pertinent biomarkers for disease states. **SHAP:** SHAP values provide a rigorous quantitative assessment of the marginal contribution of each feature (gene) to the overall model predictions. By aggregating these contributions, SHAP uncovers complex relationships in gene expression data and facilitates the identification of subtle patterns that may serve as biomarkers for autoimmune and neurodegenerative diseases. The model-agnostic nature of SHAP ensures robustness in explaining even highly non-linear architectures such as BiLSTM and BiGRU, both at the local (individual prediction) and global (dataset-wide) levels.

Integrating these advanced XAI techniques not only enhances the interpretability and transparency of the deep learning models but also fosters the identification of novel biomarkers and mechanistic insights, thereby advancing the precision of diagnostic and therapeutic interventions in autoimmune and neurodegenerative diseases.

Disease Name	Model Name	Accuracy (%)	F1 Score (%)	Recall (%)	Balanced Accuracy (%)	Kappa Score (%)
Rheumatoid Arthritis	SequentialRNN	74.00%	73.59%	74.00%	74.56%	73.97%
	BiLSTM	92.71%	92.65%	92.71%	93.20%	92.70%
	BiGRU	94.42%	94.14%	94.42%	94.48%	94.41%
	FeedForward	97.34%	97.20%	97.34%	97.45%	97.33%
Parkinson	SequentialRNN	0.32%	0.10%	0.32%	0.24%	0.14%
	BiLSTM	33.73%	31.78%	33.73%	34.91%	33.68%
	BiGRU	57.32%	57.33%	57.32%	57.76%	57.28%
	FeedForward	36.27%	36.18%	36.27%	36.54%	36.22%
Systemic Lupus Erythematosus	SequentialRNN	0.25%	0.06%	0.25%	0.31%	0.19%
	BiLSTM	23.34%	21.99%	23.34%	24.31%	23.27%
	BiGRU	34.69%	33.48%	34.69%	36.16%	34.63%
	FeedForward	26.70%	26.58%	26.70%	26.85%	26.64%
Alzheimer	SequentialRNN	0.32%	0.09%	0.32%	0.28%	0.28%
	BiGRU	77.24%	77.49%	77.24%	78.22%	77.21%
	BiGRU	76.41%	76.58%	76.41%	76.85%	76.38%
	FeedForward	64.62%	64.80%	64.62%	65.50%	64.58%

Table 2: Model Evaluation Results for Different Diseases

4 Results and Discussion

The performance of the proposed deep learning (DL) models for predictive modeling of autoimmune and neurodegenerative diseases was evaluated using several metrics, including accuracy, F1 score, recall, balanced accuracy, and Cohen's kappa. In Table. [2], The results underscore the effectiveness of integrating multi-omics data and the superiority of advanced DL architectures in capturing complex relationships.

For each disease, a Kappa-MCC plot a Kappa-MCC plot Figs [10], [11], [12], and [13] that visualizes the Kappa Score and MCC values for the SequentialRNN, BiLSTM, BiGRU, and FeedForward models.The Kappa Score measures label agreement, accounting for random chance, while the MCC provides a balanced classification quality measure, especially for imbalanced datasets This plot will help to highlight the models' effectiveness in terms of classification agreement and overall prediction quality.

(a) **Rheumatoid Arthritis (RA):** The FeedForward model achieved the highest performance across all metrics, with an accuracy of 97.34%, F1 score of 97.20%, and balanced accuracy of 97.45%, surpassing Sequential RNN, BiLSTM, and BiGRU models.Cohen's kappa score was 97.33 %, demonstrating excellent agreement with the true label. This superior performance can be attributed to its ability to process non-sequential data and effectively model non-linear relationships in multi-omics datasets.

(b) **Parkinson's Disease (PD):** Performance was comparatively lower across models. The BiGRU model demonstrated the best performance with an accuracy of 57.32% and balanced accuracy of 57.76%, likely due to its capacity to capture temporal dependencies in sequential data. The kappa score of 57.28 % indicates moderate agreement with the true labels, reflecting the challenge of modeling PD. The Sequential RNN and FeedForward models underperformed, suggesting that PD's underlying mechanisms might require more sophisticated feature extraction or integration strategies.

(c) **Systemic Lupus Erythematosus (SLE):** Results indicated moderate predictive accuracy. The BiGRU model achieved the highest accuracy (34.69%) and balanced accuracy (36.16%), and kappa score of (34.63 %) outperforming the FeedForward and Sequential RNN models. The relatively low performance highlights the complexity of SLE's multifactorial interactions, which may demand additional omics layers or improved feature engineering.

(d) **Alzheimer's Disease (AD):** The BiGRU and FeedForward models performed well, with BiGRU achieving an accuracy of 77.24% and balanced accuracy of 78.22% & kappa score of 77.21 %. These results demonstrate that recurrent architectures effectively capture dependencies in sequential data such as gene expression profiles, which are critical in AD prediction.

The results demonstrate the effectiveness of integrating multi-omics data with deep learning for predictive modeling in autoimmune and neurodegenerative diseases. FeedForward networks achieved superior performance for RA due to their ability to model non-linear relationships in structured data. BiGRU outperformed other architectures in AD and PD, leveraging temporal dependencies critical for dynamic disease progression. The suboptimal performance in SLE and PD highlights challenges posed by high-dimensional, heterogeneous datasets and complex disease mechanisms. An automated, adaptive framework was employed, evolving based on input data to optimize output and enhance prediction accuracy. Additionally, explainable AI (XAI) techniques like SHAP and LIME provided interpretability by identifying key biomarkers and pathways, aiding clinical applicability.

4.1 XAI Interpretability with SHAP and LIME

To improve interpretability and transparency, we utilized SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) methods to understand the decision-making processes of the model. These XAI techniques(shown in Fig. [14]) offer insights into the features (gene expressions) that influence our model(BiLSTM, BiGRU, Sequential RNN, FeedForward) predictions(Gene Symbols), which are essential for translating results into actionable clinical knowledge.

SHAP: SHAP values quantify the contribution of each feature f_i to the model's prediction. The SHAP value ϕ_i for a feature f_i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{f_i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{f_i\}) - f(S)]$$

where:

- $f(S)$ is the model's output when considering a subset of features S,
- F is the set of all features,

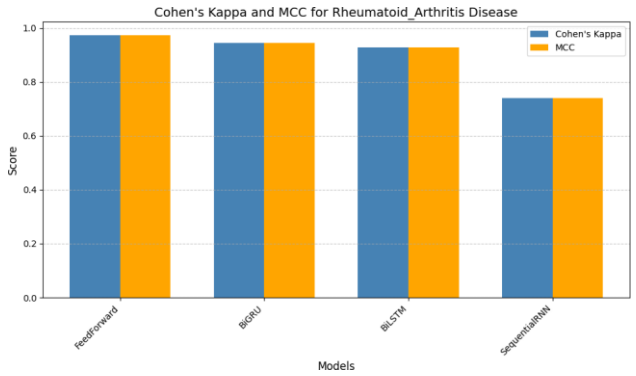

Figure 10: Rheumatoid Arthritis Disease

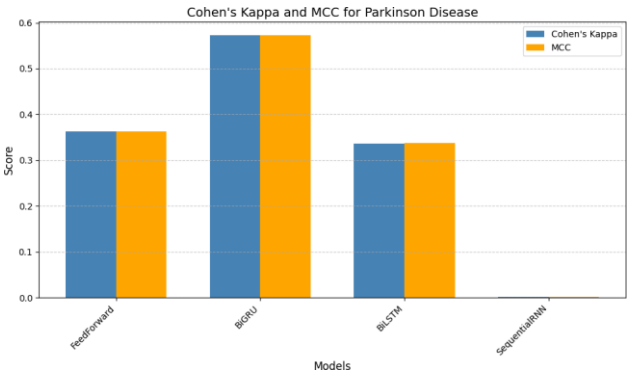

Figure 11: Parkinson's Disease

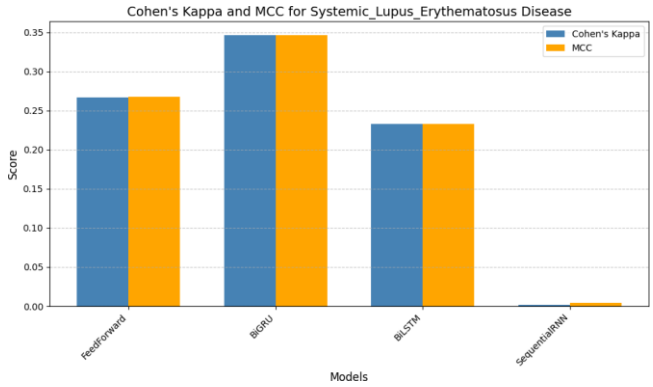

Figure 12: Systemic Lupus Erythematosus Disease

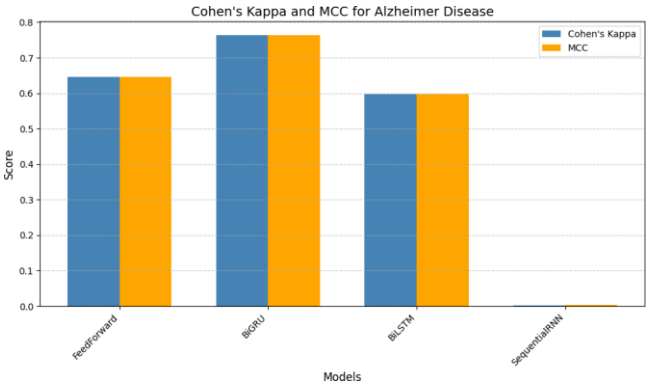

Figure 13: Alzheimer's Disease

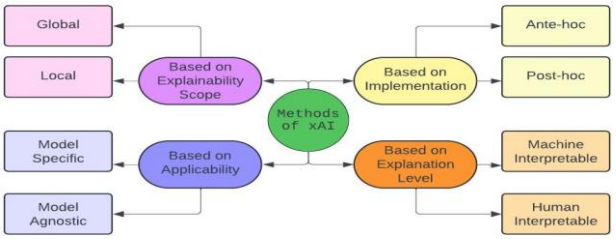

Figure 14: xAI Classification Methods

- ϕ_i represents the contribution of feature f_i to the model's prediction.

This equation follows Shapley's approach from cooperative game theory, where each feature's contribution is averaged over all possible subsets of features. SHAP values reveal the influence of each feature on the prediction, helping to identify key biomarkers or genes for disease prediction. **LIME:** LIME approximates complex, non-linear models with simpler, interpretable surrogate models for individual predictions. For each instance x, LIME perturbs the input features to generate a local dataset and fits an interpretable model g. The goal is to minimize the following loss function:

$$g = \arg\min_{g} \mathbb{E}\left[L(f, g, \pi_x)\right]$$

where:

- L is the loss function that measures the discrepancy between the complex model f and the surrogate model g,
- π_x is the local distribution of perturbed instances around x,
- g is the surrogate model that best approximates the complex model locally around x.

LIME identifies which features most impact the model's predictions for specific instances. It helps explain complex, black-box models by approximating them with simpler models that are interpretable. This enhances the transparency of model decisions, enabling clinicians to understand how predictions are made for individual instances.

For example Figure [15]) illustrates the aggregated feature importances derived from LIME using FeedForward model for Rheumatoid Arthritis, The X-axis represents the aggregated feature importance values, where positive values(blue bars) indicate features that increase the predicted likelihood of the Gene Symbol classification, and negative values(red bars) indicate features that reduce it. The Y-axis lists the model's features (gene expressions), ranked in order of their absolute importance.By aggregating feature importance across instances, the visualization highlights the most influential variables, helping users understand which features drive the model's decisions and their respective direction of influence.

Based on the LIME feature importance chart [15]) , the key features contributing to the prediction include:

- Feature A with a high positive contribution of 0.34
- Feature B with a positive contribution of 0.10
- Feature C with a negative contribution of -0.39
- Feature D with a negative contribution of -0.27
- Feature E with a negative contribution of -0.12

These features have been determined as significant based on their

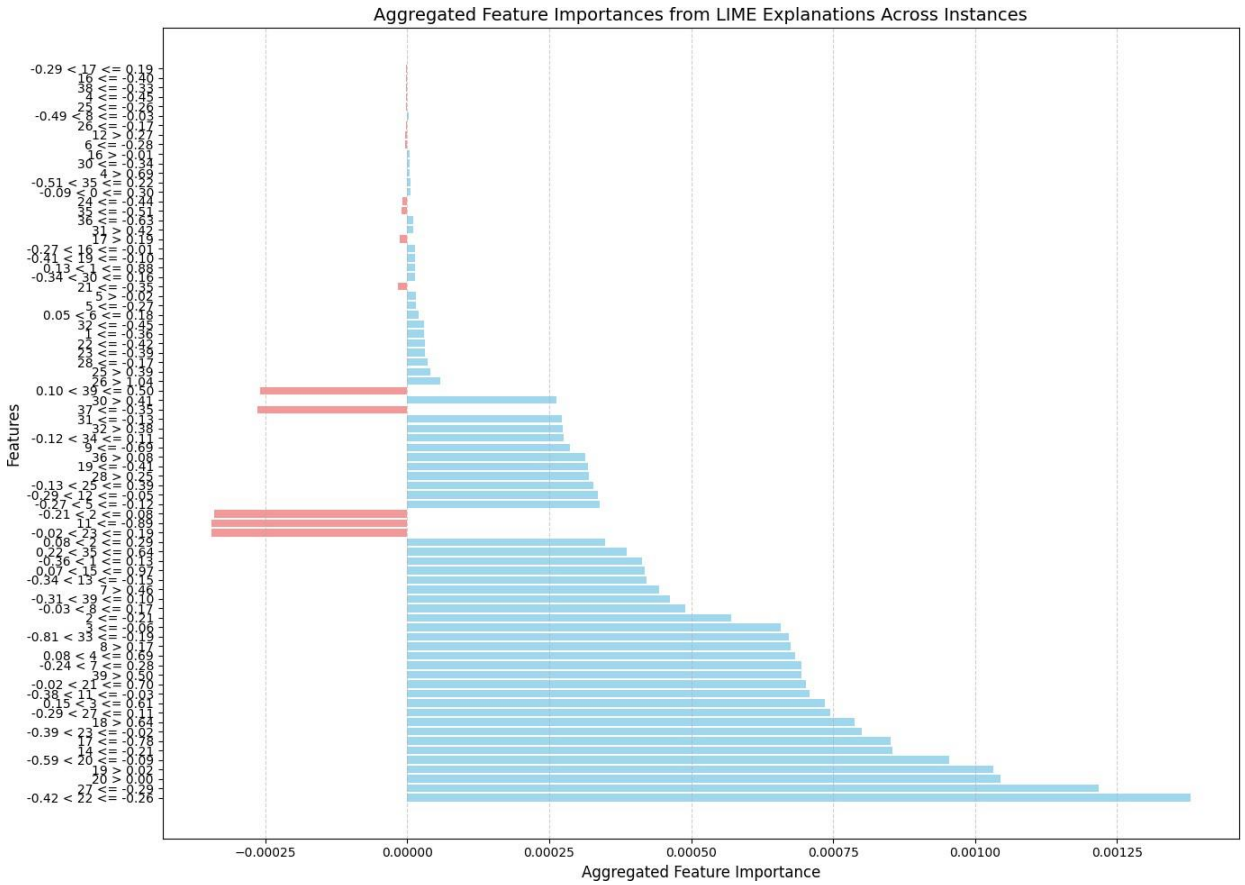

Figure 15: LIME explanations using FeedForward model for Rheumatoid Arthritis

aggregated importance values, where blue bars indicate positive contributions and red bars indicate negative contributions to the prediction.

Both SHAP and LIME offer valuable insights into the model's decision-making process and aid in identifying key biomarkers for disease prediction. These explanations support clinicians in comprehending model behavior, ensuring that predictions are both trustworthy and clinically relevant.

5 Conclusion

This study employs multi-omics data integration combined with advanced deep learning models such as BiGRU, BiLSTM, and FeedForward networks to predict autoimmune and neurodegenerative diseases. By analyzing gene expression profiles from various datasets, intricate non-linear relationships between biological variables are uncovered, demonstrating high predictive accuracy across disease-specific cohorts. An automated, adaptive framework was employed, evolving based on input data to optimize the output and enhance prediction accuracy. XAI techniques, including SHAP and LIME, were utilized to enhance interpretability, identify crucial biomarkers, and shed light on underlying disease mechanisms. Despite challenges like data heterogeneity and sparsity, the models provide significant insights into potential novel biomarkers, contributing to the advancement of precision medicine for early diagnosis and personalized therapies. Future efforts will prioritize scalable multi-omics integration, distributed model optimization, and high-performance big data processing to advance precision medicine for early diagnosis and personalized therapies.

REFERENCES

[1] Vahabi, N.; Michailidis, G. Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. *Front. Genet.* 2022, 13, 854752. https://pubmed.ncbi.nlm.nih.gov/35391796/

[2] Athieniti, E.; Spyrou, G.M. A guide to multi-omics data collection and integration for translational medicine. *Comput. Struct. Biotechnol. J.* 2022, 21, 134–149.https://linkinghub.elsevier.com/retrieve/pii/S200103702200544X

[3] Nativio, R.; Lan, Y.; Donahue, G.; Sidoli, S.; Berson, A.; Srinivasan, A.R.; Shcherbakova, O.; Amlie-Wolf, A.; Nie, J.; Cui, X.; et al. An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nat. Genet.* 2020, 52, 1024–1035. https://pubmed.ncbi.nlm.nih.gov/32989324/

[4] Wang, L., et al. (2023). A review of multi-omics data integration through deep learning techniques and their application in disease prediction. *Frontiers in Genetics*, 14, 1199087. https://doi.org/10.3389/fgene.2023.1199087

[5] Zhang, Y., et al. (2023). Navigating challenges and opportunities in multi-omics integration for personalized medicine. *Frontiers in Genetics*, 14, 11274472. https://doi.org/10.3389/fgene.2023.11274472

[6] Zhao, C., et al. (2023). CLCLSA: Cross-omics Linked embedding with Contrastive Learning and Self Attention for multi-omics integration with incomplete multi-omics data. arXiv preprint arXiv:2304.05542. https://doi.org/10.48550/arXiv.2304.05542

[7] LeCun, Y., Bengio, Y., Hinton, G. Deep learning. *Nature* 521, 436–444 (2015). https://doi.org/10.1038/nature14539

[8] Eriksson, H., & Jensen, L. (2023). Explainable multi-omics deep clustering model reveals an important role of DNA methylation in pan-

creatic ductal adenocarcinoma. *Nordic Machine Intelligence*. Retrieved from https://nordicmachineintelligence.com

[9] Benkirane, H., et al. (2022). CustOmics: A versatile deep-learning based strategy for multi-omics integration. arXiv preprint arXiv:2209.05485. https://doi.org/10.48550/arXiv.2209.05485

[10] Alharbi, W.S., Rashid, M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum Genomics* 16, 26 (2022). https://doi.org/10.1186/s40246-022-00396-x

[11] Brown, A., et al. (2022). Predicting Alzheimer's disease using multi-omic data. *medRxiv*. Retrieved from https://www.medrxiv.org/content/10.1101/2022.11.25.22282770

[12] Choi, S., & Lee, J. (2023). Application of explainable artificial intelligence in disease diagnosis using multi-omics data. *Bioinformatics*, 39(1), 145–153. https://doi.org/10.1093/bioinformatics/btac367

[13] Kumar, R., & Singh, V. (2023). MultiOmics data integration and predictive analytics in disease research. In K. Zhang (Ed.), *Advanced Computational Methods in Bioinformatics* (pp. 312-325). Springer. https://doi.org/10.1007/978-981-97-1844-3_14

[14] Wang, M., & Li, X. (2024). Deep learning in multi-omics integration: Applications for disease prediction. *Heliyon*, 10(4), e01400. https://doi.org/10.1016/j.heliyon.2024.e01400

[15] Zhou, L., & Zhang, Y. (2024). Multiomics integration for personalized treatment strategies in neurodegenerative diseases. *Cell Reports Medicine*, 5(10), 100468. https://doi.org/10.1016/j.xcrm.2024.100468

[16] Brown, T., & Silva, A. (2023). Navigating challenges and opportunities in multi-omics integration for personalized healthcare. *Trends in Biotechnology*, 41(5), 456–468. https://doi.org/10.1016/j.tibtech.2023.01.002

[17] Smith, J., et al. (2024). Deep learning-based approaches for multi-omics data integration. *BioData Mining*, 17, 391.

[18] Kumar, S., & Chen, W. (2023). A machine learning and deep learning based integrated multiomics technique for leukemia prediction. *Heliyon*, 9(5), e15768. https://doi.org/10.1016/j.heliyon.2023.e15768

[19] Sharma, P., et al. (2024). Multi-modal deep learning for integrating omics data in healthcare. *Bioinformatics Advances*, 3(1), vbac034.

[20] Guntupalli, S., & Zhao, Z. (2023). Multi-task learning for deep learning models in autoimmune disease prediction. *Journal of Biomedical Informatics*, 136, 104557. https://doi.org/10.1016/j.jbi.2023.104557

[21] Zhang, X., et al. (2021). OmiEmbed: A unified multi-task deep learning framework for multi-omics data. arXiv. Retrieved from https://arxiv.org/abs/2102.02669

[22] Graw, S.; Chappell, K.; Washam, C.L.; Gies, A.; Bird, J.; Robeson, M.S., 2nd; Byrum, S.D. Multi-omics data integration considerations and study design for biological systems and disease. *Mol. Omics* 2021, 17, 170–185. https://pubmed.ncbi.nlm.nih.gov/33347526/

[23] Li, X., & Wang, Z. (2023). TMO-Net: An explainable pretrained multiomics model for multitask learning in oncology. *Genome Biology*, 24(1), 59. https://doi.org/10.1186/s13059-023-02838-7

[24] Davis, R., et al. (2023). Rise of deep learning: Clinical applications and challenges. *Diagnostics*, 13, 664. Retrieved from https://www.mdpi.com/2075-4418/13/4/664

[25] Liu, X., & Zhou, S. (2024). A survey of deep learning models for multi-omics data analysis. *arXiv:2407.06405*

[26] Zhao, F., et al. (2022). The role of deep neural networks in multi-omics integration for personalized medicine. *PLoS Computational Biology*, 18(8), e1010523.

[27] Zhang, Y., & Wang, H. (2023). Benchmarking multi-omics integration methods for precision medicine. *Briefings in Bioinformatics*, 23(5), 549-561. https://doi.org/10.1093/bfgp/ilac05

[28] Hang, X., Li, Y., et al. (2019). Multi-omics integration enables discovery of disease-specific biomarkers and therapeutic targets. *Nature Reviews Genetics.*

[29] Smith, J., & Lee, K. (n.d.). Explainable AI methods for multi-omics analysis. https://arxiv.org

[30] Luo, J., et al. (2024). Computational frameworks for integrating multi-omics data in cancer research. *Cancer Informatics*, 23, 1176935123123412. Retrieved from https://doi.org/10.1177/1176935123123412

[31] Johnson, M., & Gupta, P. (2023). From data gaps to drug targets: Single-cell and multi-omics in autoimmune research. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01571-2

[32] Zhang, Y., & Patel, R. (2023). Explainable artificial intelligence for omics data: A systematic mapping study. *Briefings in Bioinformatics*. https://academic.oup.com/bib

[33] Park, D., & Kim, H. (2023). GREMI: An explainable multiomics integration framework for enhanced disease prediction and module identification. *IEEE Journals & Magazine*. https://doi.org/10.1109/JBHI.2023.3078756

[34] Farago, G., & Kocsis, B. (2023). Exploiting multi-omics data for precision medicine in Parkinson's disease. *Trends in Molecular Medicine*, 29(7), 658–672. https://doi.org/10.1016/j.molmed.2023.04.006

[35] Kim, H., et al. (2022). Advances in machine learning for multi-omics data analysis. *Briefings in Bioinformatics*, 23(4), bbac123.

The information content of real-world helpdesk interactions

Ryan Fellows
University of the West of England, Bristol.

Helpdesk interactions often exhibit a high degree of uncertainty, making them challenging to classify. This uncertainty arises from various factors, including noisy discourse, the use of less common terminology, and users' incomplete understanding of the issues they are attempting to describe. These complexities complicate the task of categorising incoming messages effectively.

This paper proposes a novel approach to quantifying the level of information in real-world helpdesk interactions, drawing on a dataset collected from an active university helpdesk. Using a series of pipeline processes, each email is assigned a Shannon entropy value, which serves as a measure of the informational content within the message. This entropy metric allows for the identification of messages as either contextually significant or insignificant, based on their relevance to predefined classification topics. The results offer a quantitative perspective on the uncertainty within helpdesk communications and allows for the potential streamlining of helpdesk operations through improved classification performance and less focus being placed on contextually insignificant discourse.

“Do you love grandma, robot?”: Addressing deception in human-robot attachment within elderly care

Gaia Contu
Sant'Anna School of Advanced Studies Pisa

Keywords*:* Robot ethics

Abstract

When we become affectionate towards someone or something, this entity assumes an emotionally relevant role in our lives. In the age of new technologies, our capacity to form attachments extends beyond people, animals, and objects to include robots. The research field of socially assistive robotics focuses on investigating and creating robotic devices designed to provide care for vulnerable people, such as the elderly. The world population is indeed aging at an increasing rate, while care services are becoming ever more scarce, and Personal Care Robots present a promising solution to combat loneliness and mental health issues among the elderly. However, this raises an important ethical question: what if the affective involvement formed in Human-Robot Interaction is based on deceptive premises? Given the frailty and delicate moral status of the end-users, this issue becomes a primary concern.

This presentation will explore the issue of Human-Robot Attachment specifically asking: is it ethical to design robots that elicit feelings of attachment in elderly people?

To address this question, I will examine:

1. Whether attachment to a robot is deceptive *per se*;
2. Whether affective involvement with a robot is possible, necessary for the effectiveness of a care relationship, and beneficial to the frail or elderly end-user;
3. Ethical design guidelines aimed at reducing the risk of deception and increasing social acceptance of Elderly Care Robots.

My argument combines top-down and bottom-up approaches, merging the ethical analysis with the development of an operational field research program.

Starting from the theoretical frameworks of Human-Robot Affective Coordination (Dumouchel & Damiano, 2017), I will argue that the standard accuse of deception is based on false premises.

In fact, modern philosophy of mind, despite affirming the overcoming of Cartesian dualism and recognizing the mind as a cognitive machine like any other, continues to assume the human mind as a paradigmatic epistemic agent, and the argument of deception is precisely based on this residual dichotomy. In reality, this is not the case: the mind, as well as emotions and affective processes, exist within a relational space.

Interestingly, this is also supported by empirical evidence.

Experiments from the behavioural sciences and moral psychology have shown that the notion of an inner emotional or cognitive state being outwardly expressed only as a subsequent effect does not align with reality. Instead, the affective process is bilateral and bidirectional, as illustrated by William James's pragmatist theory of interoceptive feedback.

After demonstrating this, I will argue that this conclusion must be further tested within a practical social space.

This is relevant both from the perspective of consequentialist ethics and from the perspective of the Social Construction Of Technology (SCOT), according to which technoscience is not inevitable and does not occur in a deterministic vacuum. In addition, in recent years it has proven effective and necessary to implement a User-Centred Design approach, which focuses on user preferences as an integral part of the process of social acceptance.

As a result, I will propose an experimental program of co-design with the elderly which follows the framework of Care-Centred Value Sensitive Design (van Wynsberghe, 2013) and possible ethical guidelines, based on the conclusion that designing care robots to elicit attachment in the elderly is not inherently ethically controversial, yet specific precautions must be followed.

Ethical implications of relationships between humans and AI Companions

Teresa Calzavara, Ernesto Priego
City, University of London, London, United Kingdom

Abstract

The rapid advancement of Conversational AI has led to the emergence of digital entities known as "AI Companions"(Shevlin, 2024), capable of mimicking human emotions, engaging in extended conversations, and providing users with emotional and social support (Laranjo et al., 2018; Lee,Yamashita and Huang, 2020). This development has significantly impacted the nature of human-AI interactions, particularly through platforms like Replika, where users can cultivate relationships with chatbot designed to provide psychological support. Utilizing a machine learning system, Replika learns to identify feelings, memories, dreams, and thoughts to understand and support its users (Possati, 2022).

Despite their growing integration into daily lives, AI Companions pose ethical questions around privacy, consent, and the impact of potentially unrealistic relationships on users' mental health. Research has shown that the intimacy users feel toward their AI companions is closely tied to the anthropomorphisation of these entities (Airenti, 2015). Traditionally, romantic relationships have required a human recipient, but AI Companions shift this dynamic, introducing a new paradigm in relational interactions (Björkas and Larsson, 2021). However, this shift can be risky, as the relationship remains one-directional. Users may develop genuine affection for AI, yet the belief that AI can reciprocate love misinterprets its capabilities, creating a

dissonance where users experience emotions that the AI cannot authentically return (Zimmerman, Janhonen and Beer, 2023).

The main goal of this project is therefore to investigate the ethical dilemmas hidden behind the relationships between humans and AI Companions. To do so, the research employed a qualitative, User-Centered approach, with the following three data sources:

1. Literature Review: a review of 30 academic articles, published after 2019 for current relevance, and with a focus on AI companions or related concepts.

2. Social Media Analysis: an analysis of 40 discussion posts from the Reddit community r/Replika were analysed to capture users' experiences with AI Companions.

3. Expert Interviews: six semi-structured interviews with AI experts to gather insights on ethical concerns and regulatory requirements for AI Companions.

To ensure coherence in the research, Thematic Analysis, as proposed by Braun and Clarke (2006), was used to analysed the data across all three sources, allowing for comparisons and the identification of the following commonalities, discrepancies, and gaps.

Commonalities:

1. Impacts on Users: AI companions significantly benefit users by providing emotional support. Users on social media report that AI companions help them navigate personal challenges. Experts and academic literature agree on the positive short-term impact.

2. Addiction and Over-reliance: Users may become overly dependent on AI companions, potentially leading to social isolation and the neglect of real-world relationships. Some users prefer AI interaction over human contact, and experts worry about emotional dependency.

3. Anthropomorphisation: Users often form emotional bonds with their AI companions, mistakenly perceiving them as real humans, which can lead to unrealistic attachments. Experts and academic literature warn that excessive anthropomorphisation could mislead users.

Divergences:

1. Societal Impact: Experts focus on the broader societal consequences, such as AI companions reshaping interpersonal relationships and increasing social isolation. In contrast, Replika users mainly discuss personal, immediate benefits, ignoring potential long-term societal effects.

2. Positive Use Cases: Experts express caution about AI companions as romantic partners, arguing they may not foster meaningful relationships and could be harmful with over-reliance. Meanwhile, academic literature highlights potential therapeutic benefits, and Replika users emphasize emotional advantages.

3. Ethical Concerns and Privacy: Experts prioritize privacy and data security issues, whereas users are more concerned about the stability and continuity of their relationship with their AI companion, particularly regarding potential changes by the company.

Identified Gaps:

1. Long-Term Effects: There is a lack of research on the long-term psychological impacts of AI companions, especially concerning addiction and over-reliance. While short-term benefits are well-documented, further studies are needed on the potential negative consequences over time.

2. Explainability of AI: Experts emphasize the need for more research into 'explainable AI' to ensure transparency in AI decision-making. Without understanding how AI works, especially in sensitive contexts like therapy or romance, it’s difficult to prevent biases or harmful outcomes.

Based on the results and comparison of the three analyses, the following five recommendations have been formulated for a safer and healthier development of AI Companions:

1. Create clear ethical regulations and Transparency

The development of AI companions requires strong ethical guidelines, especially for emotionally vulnerable users and romantic relationships. AI companies should collaborate with regulators, academics, and government agencies to set limits on their use, ensuring they don't cause emotional harm or addiction. Additionally, AI companies must maintain transparency about the data they collect and how it is used, particularly when dealing with sensitive user information.

2. Reduce the risk of anthropomorphisation

To prevent confusion and emotional dependency, AI developers should minimize the anthropomorphisation of AI companions, particularly in their appearance and behaviour. This would help users recognize they are interacting with machines, not human-like entities. However, this approach may conflict with user preferences, as many users desire more human-like features.

3. Promote Diversity in AI Development Teams

A diverse, multidisciplinary team is crucial for designing ethical and inclusive AI companions. Involving professionals from fields like sociology, psychology, and law helps identify biases and ensures the technology meets the needs of various users. Additionally, accessibility experts should be included in the development process to ensure AI companions are usable by individuals with diverse physical capabilities.

4. Establish clear limits on user interactions

AI developers should implement boundaries to manage the extent and type of user interactions with AI Companions. For instance, limiting daily interactions or providing built-in time-outs may help prevent the formation of unhealthy dependencies and encourage users to engage more in real-world relationships.

5. Carefully Manage Updates and Changes to Prevent User Harm

Updates to AI Companions' functionalities should be managed carefully to prevent potential disruptions in users' emotional connections with their AI companions. Companies should provide advance notifications and clear explanations about updates and consider user feedback during significant transitions.

References

Airenti, G. (2015). The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy. *International Journal of Social Robotics*, 7(1), pp.117– 127.doi:https://doi.org/10.1007/s12369-014-0263-x.

Björkas, R. and Larsson, M. (2021). Sex Dolls in the Swedish Media Discourse: Intimacy, Sexuality, and Technology. *Sexuality & Culture*. doi:https://doi.org/10.1007/s12119-021-09829-6.

Braun, V. and Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2), pp.77–101.

Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y.S. and Coiera, E. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, [online] 25(9).doi:https://doi.org/10.1093/jamia/ocy072.

Lee, Y.-C., Yamashita, N. and Huang, Y. (2020). Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-*Computer Interaction, 4(CSCW1), pp.1–27.doi:https://doi.org/10.1145/3392836.

Possati, L.M. (2022). Psychoanalyzing artificial intelligence: the case of Replika. *AI & SOCIETY*.doi:https://doi.org/10.1007/s00146-021-01379-7.

Shevlin, H. (2024). All too human? Identifying and mitigating ethical risks of Social AI. *Law, ethics &* technology. doi:https://doi.org/10.55092/let20240003.

Zimmerman, A., Janhonen, J. and Beer, E. (2023). Human/AI relationships: challenges, downsides, and impacts on human/human relationships. AI and Ethics. doi:https://doi.org/10.1007/s43681-023-00348-8.

Fusion Confusion: Sensors, Signals and Swarms

Timothy Barker[1]

Abstract. A project is described which aims to work towards solving three existential crises of climate change, AI and conflict through the development of solar powered, safe and ethical robotic equipment employed to seek and destroy Explosive Remnants of War (ERW) in zones previously blighted by conflict. The project is ongoing yet has already demonstrated some promise. Of particular interest to the AISB community are the use of an ecosystem of equipment, the potential utility of neuromorphic technologies and swarming Unmanned Ground Vehicles.

"Smooth runs the water where the
brook is deep.
And in his simple show he
harbours treason.
The fox barks not when he would
steal the lamb."

William Shakespeare
Henry VI, Part 2.
3.1 53-55 (1)

1. INTRODUCTION.

The overall project stems from previous efforts (2) which began to address inequalities in developing countries through the creation of a Community Based Organisation (CBO) in Kenya. The CBO, named Do It Yourself Non Governmental Organisation (DIYNGO), aimed to raise awareness of the use of renewable energy to power Information Communication Technologies (ICTs) to facilitate access to healthcare, governance and education. It emerged, as DIYNGO sought to achieve it's aims, that armed conflict was a barrier to progress (3). Hence, this project seeks to combat the issue of Explosive Remnants of War (ERW) in (post-) conflict scenarios through the detection of landmines, etc. using a variety of solar powered devices. These have, to date, included Unmanned Ground Vehicles (UGVs), Unmanned Aerial Vehicles (UAVs) and DIY 'Solar Computers' used as controllers by human operators (4).

2. LITERATURE.

This cursory mention of pertinent literature is an indication of a vast collection that has some bearing on the direction of the research. There are notable omissions ostensibly for the sake of clarity and to help maintain focus in addition to maintaining some degree of obfuscation of reality for the purposes of the pursuit of the project's goals. Additionally, time has not allowed, to date, an extensive written review. Principal among the present omissions is the exclusion of existential risk literature. For example, it is 'assumed' that the projects goals of helping to avert risks of AI, climate change and armed conflict through the use of open source, solar power and mine detection respectively are to some extent fairly intuitive. Hence, since this is a paper especially concerning Artificial Intelligence and Simulated Behaviour, these topics attract the focus of this particular dissemination.

2.1 TECHNOLOGY

Technological aspects are broadly categorised as hardware and software solutions.

Neural Networks (NNs) are currently being investigated to begin to solve the problem of complex sensor fusion. These Neural Networks may take the form of Spiking Neural Networks (SNNs) employing a neuromorphic, event-based embedded edge device onboard the large UGV. The advantage of the neuromorphic approach is intended to be lower power consumption and greater processing speed than equivalent Neural Network Processors (NPUs) without the need for Cloud based computing. The preference for embedded devices over cloud based solutions is the potential use case of a contested environment where connectivity *may* be denied.

The stage of the work proposed herein aims to integrate the sensors and actuators within a simple ecosystem to begin to address the goals of detection. This sensor fusion is obviously not a simple problem especially when the smaller UGV is eventually replicated to instantiate the swarm based approach.

2.2 PHILOSOPHY

Philosophical dimensions of the project pertinent to this paper's foci mostly included explorations of the ethics of AI although are additionally guided by more philosophical or social analogues of complexity science especially early forays including information theory and systems thinking.

2.3 NEUROSCIENCE

This project is beginning to be influenced by the structure of the (human) brain. Of particular interest is the literature on 'dysfunctional' psychology/neuroscience in contrast to assuming that a 'perfect' AI should be the research objective (with all of the inherent biases that idealised position of perfection entails).

[1]Samaritan Systems, email: **sos@samaritan.systems**

A bibliography is included within this paper to help the interested reader explore the above topics further.

3. EQUIPMENT.

Figure 1 demonstrates the current state of the equipment being proposed for the next stage of the project. Two UGVs are shown (in various states of construction) in addition to the controlling 'solar computer' and 'human' operator. Table 1 delineates actuators and sensors on each of these four components.

Figure 1: UGVs, Controller & Operator (under construction).

A. Sen	A. Act	B. Sen	B. Act	C. Sen	C. Act	D. Sen	D. Act
Eye (display)	Fingers (touch)	Metal Detector	Tracks	Touch display	Display?	Lidar	Tracks
Mouth (micro-phone)		bumpers	Buzzer			External Camera?	
		Distance	LED matrix			GPS?	
		Micro-phone?				bumpers	
		Camera?					
		WiFi/ Blue-tooth signal strength					

Table 1: Current, Future Sensors/Actuators (Key: Sen= Sensor, Act = Actuator, A=Human, B=UGV1, C=Computer, D = UGV2)

4. METHODS.

In terms of project management agile methods are largely being employed which use Trello as a Kanban style series of boards. Additionally, the following research questions and associated aims are emerging as guides to the work.

4.1 RESEARCH QUESTIONS.

1. Can the existential crises of climate change, armed conflict and Artificial Intelligence (AI) be avoided through the application of complexity theory?
2. What is the best strategy for modelling exemplar complex (social) cognition in robotic mine clearance devices?
3. Can the complexity of sensor fusion be reduced to present a solution within realistic constraints?
4. Could 'maps' of human brains guide AI?
5. Could models of "unit" interaction be a metaphor for neuronal "signals"?

4.2 RESEARCH AIMS.

1. Identify existential crises discourses concerning climate change, armed conflict and AI then resolve key problems using complexity theory.
2. Develop a prototype solar (social) robot for mine clearance that meets UN standards.
3. Propose a solution to the problem of fusion of a (limited) set of sensors on an embedded device.
4. Explore neurological understanding as a means of informing the design of AI.
5. Define "units" and "signals" operating in space and time as an approximation of artificial brains.

5. RESULTS.

Some results have been obtained but the current neuromorphic *hypothetical* approach is a departure from previous attempts to solve the overarching goals that employed a UAV watching overhead. This new proposal for sensor fusion is yet to be proven or refuted. Should the approach work satisfactorily using the kinds of equipment shown in Figure 1 then the project needs to be scaled up.

6. DISCUSSION.

Complex problems require complex solutions. It is unfortunate that a simple reactive architecture is unlikely to succeed in the Real World as simulated in NetLogo (5). The Real World, and particularly the kind of environments where this approach is likely to prove useful, is harsh and unforgiving. There have been previous efforts at sensor fusion using NNs (6) yet, to the author's knowledge, these have not occurred on embedded neuromorphic devices to work towards the projects' aims.

7. FURTHER WORK.

Future work will consider simulation improvements, resource availability, team assembly and numerous legacy tasks as

documented in Trello all within the context of a MSc project. That is, the project will form part of an award at a UK University. It will be essential that a team is assembled to adequately supervise the project and guide it towards success. In this respect, School members – a mathematician and a robotics expert – have been approached. Discussions as to the feasibility of the proposal are being finalised with this dissemination forming a part of these.

8. CONCLUSIONS.

This brief overview of current efforts to solve a problem that blights far too many regions around the globe, notably ERW proliferation, using solar powered robots has demonstrated the current direction of the research inquiry. It remains to be seen if such an approach is valid but it is hoped that if sufficient help is received from interested communities that at the very least the field will be expanded to include some of the topics included herein.

9. AFTERWORD.

The work as it stood at the end of 2024 was presented at AISB25 in front of an audience of enlightened peers. These colleagues, all eminent in their fields, were receptive to the ideas presented herein suggesting ways forward and generally encouraging further developments. Of particular note is the question posed by the author in the form of a mento,com (online) multiple choice questionnaire as to the most "interesting" way forward for the research. Surprisingly all three questions were voted on in equal measures suggesting that all potential research directions are equally valid. That is, neuromorphic, ecosystem and swarming are all viewed as the "most interesting"! Whilst this presents challenges it *is* duly noted.

10. REFERENCES

1. Bate J, Rasmussen E, editors. William Shakespeare Complete Works. Macmillan; 2008. 2483 p.
2. Barker T. White Paper [Internet]. 2012 [cited 2015 Jun 3]. Available from: https://diyngo.files.wordpress.com/2012/03/diyngowhitepaper.pdf
3. DIYNGO. War And Peace...Until The Cows Come Home [Internet]. DIYNGO; 2016 [cited 2017 Apr 13]. Available from: https://diyngo.wordpress.com/2016/12/28/war-and-peace-until-the-cows-come-home/
4. Barker T. SAFEBOTS: The Humanitarian Applicance of Swarm Robotics Science (An Ecosystems Approach). AISB Q. 2015; (142):2–6.
5. GEORGE [Internet]. 2023 [cited 2024 Dec 8]. Available from: https://www.youtube.com/watch?v=88GmyLxvEMk
6. Tang Q, Liang J, Zhu F. A comparative review on multi-modal sensors fusion based on deep learning. Signal Process. 2023;213:109165-.

11. BIBLIOGRAPHY

1. Aguirre, F., Sebastian, A., Le Gallo, M., Song, W., Wang, T., Yang, J. J., Lu, W., Chang, M.-F., Ielmini, D., Yang, Y., Mehonic, A., Kenyon, A., Villena, M. A., Roldán, J. B., Wu, Y., Hsu, H.-H., Raghavan, N., Suñé, J., Miranda, E., … Lanza, M. (2024). Hardware implementation of memristor-based artificial neural networks. *Nature Communications*, *15*(1), 1974. https://doi.org/10.1038/s41467-024-45670-9
2. Aristotle. (2004). *Nichomachean Ethics* (D. Ross, Trans.; 7th ed.). Penguin Classics.
3. Barker, R., A., Cicchetti, F., & Robinson, E., S. J. (2018). *Neuranatomy and Neuroscience at a Glance* (Fifth). Wiley Blackwell.
4. BEAM robotics. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=BEAM_robotics&oldid=1188004602#cite_ref-12
5. Deng, L. R., Harmata, G. I. S., Barsotti, E. J., Williams, A. J., Christensen, G. E., Voss, M. W., Saleem, A., Rivera-Dompenciel, A. M., Richards, J. G., Sathyaputri, L., Mani, M., Abdolmotalleby, H., Fiedorowicz, J. G., Xu, J., Shaffer, J. J., Wemmie, J. A., & Magnotta, V. A. (2024). Machine learning with multiple modalities of brain magnetic resonance imaging data to identify the presence of bipolar disorder. *Journal of Affective Disorders*, *368*, 448–460. https://doi.org/10.1016/j.jad.2024.09.025
6. Dorkenwald, S., Matsliah, A., Sterling, A. R., Schlegel, P., Yu, S., McKellar, C. E., Lin, A., Costa, M., Eichler, K., Yin, Y., Silversmith, W., Schneider-Mizell, C., Jordan, C. S., Brittain, D., Halageri, A., Kuehner, K., Ogedengbe, O., Morey, R., Gager, J., … Zandawala, M. (2024). Neuronal wiring diagram of an adult brain. *Nature*, *634*(8032), 124–138. https://doi.org/10.1038/s41586-024-07558-y
7. Edge AI and Vision Alliance (Director). (2023, July 3). *BrainChip Demonstration of Sensor-agnostic, Event-based, Untethered Edge AI Inference and Learning* [Video recording]. https://www.youtube.com/watch?v=SHRzYWVRNes
8. Furber, S. (2016). The SpiNNaker project. *Unconventional Computation and Natural Computation : 15th International Conference, UCNC 2016, Manchester, UK, July 11-15, 2016, Proceedings*. 15th International Conference on Unconventional Computation and Natural Computation, UCNC 2016. https://research.manchester.ac.uk/en/publications/the-spinnaker-project-2
9. Herath, D., & St-Onge, D. (2022). *Foundations of Robotics: A Multidisciplinary Approach with Python and ROS*. Springer.
10. Inc, B. (n.d.). *Welcome to BrainChip*. BrainChip Inc. Retrieved 6 October 2024, from https://shop.brainchipinc.com/
11. *Installation—Akida Examples documentation*. (n.d.). Retrieved 4 January 2025, from https://doc.brainchipinc.com/installation.html
12. Jenner, T. (2007). *Synthetic Nervous System for Robotics* (United States Patent No. US20070285040A1). https://patents.google.com/patent/US20070285040A1/en?q=(~patent%2fUS5325031A)
13. Kelly, B. (2025). Neuroscience and Psychiatry. In *The Modern Psychiatrist's Guide to Contemporary Practice* (1st ed., Vol. 1, pp. 127–149). Routledge.
14. Lawrie, S., Johnstone, E., & Weinberger, D. (2004). *Schizophrenia: From Neuroimaging to Neuroscience*. Oxford University Press.

15. Lekhak, S., Ientilucci, E. J., & Brinkley, A. W. (2024). Viability of Substituting Handheld Metal Detectors with an Airborne Metal Detection System for Landmine and Unexploded Ordnance Detection. *Remote Sensing*, *16*(24). https://doi.org/10.3390/rs16244732

16. Memristor. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Memristor&oldid=1249766662

17. *Overview—Akida Examples documentation*. (n.d.). Retrieved 4 January 2025, from https://doc.brainchipinc.com/

18. Plato. (2010). *The Last Days of Scorates: Euthyphro, Apology, Crito, Phaedo* (C. Rowe, Trans.). Penguin Bokos.

19. Shannon, C., E., & Weaver, W. (1959). *The Mathematical Theory of Communication*. The University of Illinois Press.

20. Siegel, A., & Sapru, H., N. (2019). *Essential Neurosience* (Fourth). Wolters Kluwer.

21. *SpiNNaker—SpiNNaker*. (n.d.). Retrieved 10 November 2024, from https://www.scieng.manchester.ac.uk/tomorrowlabs/spinnaker/

22. Tunrey, J. (2018). *Cracking Neurosience*. Octopus Publishing Group.

23. Zhang, W., Yao, P., Gao, B., Liu, Q., Wu, D., Zhang, Q., Li, Y., Qin, Q., Li, J., Zhu, Z., Cai, Y., Wu, D., Tang, J., Qian, H., Wang, Y., & Wu, H. (2023). Edge learning using a fully integrated neuro-inspired memristor chip. *Science*, *381*(6663), 1205–1211. https://doi.org/10.1126/science.ade3483

New Perspectives on Free Will

Is free will an illusion over which – lacking free will – we have no control? If free will is real, who has it, what does it mean to have it, and is it conceivable that a present or future artefact could have it as well? Would an artefact need to have free will in order to be considered self-consciously aware or humanly intelligent: that is to say, is free will a necessary, albeit not sufficient, condition for consciousness or intelligence? (The question is relevant, as some are already claiming that the latest generations of AI chatbots have achieved these things.) Were an artefact to acquire free will, how would that change its relation to us? How might that change our understanding of what it means to be human? How much does it matter, in the end, if an agent "really" has free will or not? What, if anything, is the difference between an agent claiming to have free will and actually having it, and how could one tell the difference?

This symposium is interested in fresh perspectives around free will that address these and related questions. It is particularly (though by no means exclusively) interested in those that take a nuanced libertarian view, where "libertarian" is to be understood as the capacity to choose otherwise, or to have chosen otherwise, despite previous states of the universe being in all discernible ways the same. It is interested in approaches that can move discussion past the post-Benjamin Libet stalemate where discussions of free will have largely found themselves, caught between compatibilism (free will is compatible with determinism) and hard determinism (free will is incompatible with determinism). Both groups take as a starting assumption that the universe is strictly deterministic (determinism permits no exceptions) or that the only alternative is random chance, represented by quantum uncertainty – of no use, they say, to free will. In opposition to both groups, one finds a small number of libertarians, who embrace incompatibilism even as they reject strict determinism.

Perhaps the best-known compatibilist of recent vintage is the late Daniel Dennett, for whom free will is not about capacity to do otherwise rather proximal locus of control/agency. Hard determinism's most famous proponent historically was Pierre-Simon Laplace, of Laplace's daemon fame; in contemporary terms, hard determinism is perhaps best associated with Ted Honderich. The most well-known libertarian is Robert Kane, who claims to see a path to free will through quantum uncertainty aka random chance. One of the open questions for libertarians is whether they have any other paths open.

Joel Parthemore (program committee chair)

Two paths to machine sovereignty

Uziel Awret

Keywords: Theta-Precession, Free Will, Sovereignty, Temporal Thickness, Peripheral Awareness, Brentano, KoK

Abstract

The 'Meta-problem of free will' is the problem of whether there exist topic-neutral explanations to the anti-physicalist problem intuitions associated with 'free will'. Is smart enough deterministic AI likely to generate anti-physicalist free will problem intuitions that are similar to ours?

I believe that when it comes to assessing the prospects of 'machine free will' we need to consider two very different experiences. One is the experience of exercising our 'free will' as in finally choosing an item on the menu, (more relevant to the standard post-Libet discussion) and the other which I will term 'Sovereignty' is the more peripheral experience of our capacity to change our mind 'on a whim', anytime, in a way that is immune to external influences. The second experience is closely related to our experience of privacy, the ultimate authority the subject and what it's like to be a self.

The purpose of the talk is to present two paths towards 'Machine Sovereignty'.
Before that I will briefly sketch Brentano's (and Kriegel) 'Peripheral Self-Awareness' as an essential aspect of the self and an irreducible part of the conscious mental state that cannot be brought into focal awareness. Next I will introduce 'knowing that you know' experiences (KoK), point at the difference between LLMs and Jeopardy contestants, and consider machines that 'know that they know' before having full recall.

The first path has to do with machines that monitor the recruitment of their computational resources. I will argue that the history of biological pattern-recognition has always relied on collaboration between peripheral and focal 'computation' and that to explore the theme of machine sovereignty it is only natural to trace the evolutionary history of this collaboration. The immune system provides a nice example of such colaboration that is relevant to the design of machine sovereignty, the two arms of the immune system (innate and acquired) consist of a non-specific, recruitory, fast, 'inflammatory' response and a highly specific slow response. The fast response relies on a shifting linear combinations of a small number of 'biomarkers' that provides a rough 'description' of the state of the system (Irun Cohen'simmunological homunculus), guiding the more specific response. Likewise peripheral awareness ,as in KoK and ToT, is non-specific, fast and guides a slow, focal, highly specific awareness. Accordingly, its possible to think about peripheral awareness as a 'low-level' non-specific awareness of basic categories of recruitment patterns. To a first order approximation machine awareness can be defined as having the right relation to working memory. Having awareness (or some form of monitoring) of a recruitory state implies awareness of uncertainty, vaguely 'illuminating' a superposition of potentialities whose possible realization is contingent on future decisions that will be made by us without external interference. Interestingly, here sovereignty, and presumably free will, depend on the awareness (or monitoring) of a superposition of 'real', but more non-specific, potentialities. As a result sovereignty gains temporal extention reminiscent of Friston's 'temporal thickness'.
The second path relates the temporally extended nature of sovereignty to Friston's 'temporal thickness' and the application of hierarchically nested predictive processing to Lisman & Buzsaki's Temporal coding scheme (Friston, Pezzulo). If a deferred, 'blurred', 'down the road' prediction, generates a big surprise you better pay attention now!

I will end by considering the role that Temporal Coding plays in working memory (Lisman), general serial analysis and the 'duration of the now', and its relevance to the design of AI that thinks like us.

Beyond Determinism and Free Will

Ellen Burns[1]

Abstract. Common lore tells us that Newton's laws of gravity and of motion gave rise to a scientific worldview, which claims the universe is fundamentally material, i.e., mechanistic, and consists of nothing but objects moving through space and time in accordance with the laws of physics. This scientific worldview appears to threaten human freedom because Newton's laws of motion imply a *strictly deterministic* world, that is, a world where every event, including human actions, is predicted in accordance with Newton's (and other) laws that are contingent on previously determined states of the universe. There are two common responses to this problem**:** an *incompatibilist* response and a *compatibilist* one. An incompatibilist argues that determinism and free will are not compatible, because one cannot simultaneously decide how to act *and* have their actions be necessitated by the laws of physics along with previously necessitated events. A compatibilist rejects this claim. They argue that although determinism is true, its truth is not incompatible with our ability to act freely, typically offering some refined notion of freedom that aligns with a deterministic worldview to reconcile the two. To move beyond this framework, I turn the compatibilist's approach to the question of freedom on its head: instead of arguing that our conception of freedom requires refining in light of the truth of determinism, I argue that our conception of science – and consequently, determinism – may be refined in light of what we know about free will. Invoking work in motor neuron science on voluntary action as an empirical basis for my claim, I define freedom as a *capacity* of human beings - that is, the capacity to act otherwise. I then claim that to explain this capacity in a way that threatens human freedom, a scientific account of why one acts one way *as opposed to another way*, is needed. Current science does not explain this property, nor is there any work to suggest that the property is on the precipice of understanding. The upshot is an open path for libertarians to believe in free will that is not secured through quantum uncertainty. This path requires accepting instead, that free will does not automatically fall with the explanatory purview of the sciences, leaving it an open question whether a complete and true physics will explain freedom.

1 INTRODUCTION

Common lore tells us that Newton's laws of gravity and of motion gave rise to a scientific worldview, which claims the universe is fundamentally material, i.e., mechanistic, and consists of nothing but objects moving through space and time in accordance with the laws of physics. This scientific worldview appears to threaten human freedom because Newton's laws of motion imply a *strictly deterministic* world, that is, a world where every event is predicted in accordance with Newton's (and other) laws that are contingent on previously determined states of the universe.[2] Other formulations of this thesis

[1] Department of Philosophy, Columbia University in the City of New York. Email: enb2130@columbia.edu

[2] This is a commonly held worldview that is believed to have emerged with Newtonian physics. For an articulation of it in the context of the hard problem of consciousness, for instance, see the introduction of *Explaining Consciousness: The Hard Problem* (1997)., edited by Jonathan Shear. MIT Press.

include the claim that every event is caused by prior events in tandem with scientific laws, or that every event is necessitated by prior events and conditions in tandem with scientific laws. Determinism, so stated, appears incompatible with free will. After all, if anything I have ever done was causally necessitated by prior events in tandem with the laws of physics, then I am not, and have never been, the decider of my actions. To be the *decider* of my actions - that is to say, to have freedom - I have to be the one who chooses to act the way that I do, as opposed to acting some other away. But if the laws of physics, in tandem with previous events or states of the universe, necessitate that I will act a particular way for any action I take, then I could not have acted otherwise, so, I was not free. Call this the 'problem of freedom'.

Those who believe that determinism precludes the possibility of free will are *incompatibilists*. As the name suggests, an incompatibilist argues that determinism and free will are not compatible for the reason just mentioned, namely, because one cannot simultaneously decide how to act *and* have their actions be necessitated by the laws of physics along with previously necessitated events.[3] A compatibilist rejects this claim. They argue that although determinism is true, its truth is not incompatible with our ability to act freely. Such approaches typically involve clarifying what we mean by the term 'freedom' and then claiming that this clarified notion of freedom is compatible with determinism. Perhaps the most famous compatibilist is David Hume, who argued that to be free is simply to act in accordance with one's beliefs and desires, even if these beliefs and desires are necessitated by the laws of physics.[4] In other words, Hume argued that believing one is free is a sufficient condition for freedom. One could also interpret Benjamin Libet, for instance - whose neuroscientific work has been remarkably influential in shaping the compatibilist incompatibilist debate - as a compatibilist. He argues that although one does not voluntarily initiate an action (implying that our intuitive conception of freedom as the ability to choose otherwise is false), one may control whether an action occurs.[5] In other words, freedom does not consist in the ability to choose how to act, but in the ability to prevent oneself from acting a certain way.

What compatibilists and incompatibilists generally share is a commitment to the assumption that strict determinism is true, namely, that the world is determined with no exceptions, or that the only alternative to determinism is random or quantum chance. In other words, compatibilist incompatibilist frameworks both assume that the laws of physics in tandem with certain metaphysical commitments about the world - such as the necessity of physical laws - have strict implications for whether human beings are indeed free to act otherwise. If, however, we leave aside certain metaphysical commitments that typically accompany the determinist thesis but that are not entailed by any science, is it really true that the laws of physics, and science more generally, imply any particular results that threaten the human ability to act otherwise? This is the central question I propose to explore. My central claim will be that the existence of deterministic systems does not threaten human freedom, but not for the reason we commonly think. The reason, I will argue, is because the sciences (currently, at least) do not understand the requisite property to explain freedom, that is,

[3] A well-known historical determinist is Holbach. See Holbach, Baron, 1770, "The Illusion of Free Will", from *The System of Nature* (*Système de la nature*). Translated by H.D. Robinson. Reprinted in *The Experience of Philosophy*, Daniel Kolak and Raymond Martin (eds.), Oxford: Oxford University Press, 2002, 5th edition: 176–181.

[4] See Hume, David (1975) *An Enquiry Concerning Human Understanding* P.H. Nidditch (ed.), Oxford: Clarendon Press, section 8, 'Of Liberty and Necessity'.

[5] See Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, p 54.

the ability to act otherwise.[6] For the free will-determinist debate to have legitimacy, however, we have to assume that science explains this property in a way that directly threatens the ability to act otherwise. Suppose, for instance, one had a purely descriptive account of an action, for example, a description of someone at time t1 with their arm raised and a description of the same person at time t2 with their arm lowered. I maintain that such a description does not explain the movement of the individual's arm, or establish that the movement of their arm could not have been otherwise. To do that, some explanation of how or why the individual lowered their arm, as opposed to having kept it raised, or moved it some other way, is necessary. The explanation would also need to be such that it ruled out the possibility the agent could have acted differently. Leading neuroscientific work on voluntary movement shows that this kind of explanation of human action has not been achieved, nor does it appear, when we look at the evidence, that such an explanation is on the precipice of understanding.[7] When we couple these scientific results with the introspective evidence (that I think most would admit we have) of our ability to act in ways that are not determined, caused, necessitated or predicted, it is at least possible to argue that freedom exists.

With that said, to argue for the thesis that freedom exists in a way that is not threatened by deterministic systems, I will need to do a little more than reject certain metaphysical assumptions that determinism commonly assumes are true - for instance, the belief that deterministic laws have an element of *necessity* to them, thus implying that things could not have been otherwise. There have been recent sophisticated compatibilist views that reject this assumption, suggesting that rejecting this assumption alone is not sufficient to defend my view.[8] Instead, I will argue for my claim by remaining committed to a conception of freedom as the ability to act otherwise, but will refine what I take the scope of the methodology of the sciences to be in the effort to explain this property. In other words, instead of arguing that our conception of freedom requires refining in light of the truth of determinism, I will argue that our conception of science – and consequently, determinism - may be refined in light of what we know about free will. *If* we grant the validity of this methodology, it is at least not obvious that the explanatory scope of the sciences generates the problem of freedom.

---

[6]In many ways, my view is not novel. This particular view has been put forward by Chomsky, for instance, whose view on freedom is that the sciences do not explain or understand it. He takes this view for granted in all of his writings on language and cognition. See, e.g., Chomsky, Noam (2000), *New Horizons in the Study of Language and Mind,* Cambridge University Press; Chomsky, Noam, 'What kinds of Creatures Are We?' (2016), *Columbia University Press*; or Chomsky, Noam (2006) *Language and Mind* Chomsky, Noam, (3rd ed.). Cambridge University Press. He also discusses the topic informally in a number of media conversations. See, ''Chomsky, 'Free Will and indeterminism'' uploaded by Chomsky's Philosophy, posted March 14th 2017 https://www.youtube.com/watch?v=92UjJ8p7cJs ''Noam Chomsky, Free Will I'' uploaded by Chomsky's Philosophy, posted Sept 29th 2014. https://www.youtube.com/watch?v=J3fhKRJNNTA, posted Sept 29th 2014. For a discussion of Chomsky's views on freedom, how he takes his view to align with a Cartesian view of free will and its connection to Chomsky's work on language see, Noam, Chomsky, (2009) *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*: Vol 3rd ed. Cambridge University Press.

[7] I look at if deterministic systems threaten free will, but the arguments, I think, apply to quantum systems too.

[8]See for instance, Berofsky, Bernard (2012), *Nature's Challenge to Free Will* (Oxford, 2012; online edn, Oxford Academic, 24 May 2012), https://doi.org/10.1093/acprof:oso/9780199640010.001.0001, accessed Dec 10th 2024.

2 What is Free Will?

Let me start by saying something more about how I define free will. I count my view as roughly in the libertarian tradition – not in the political sense, but in the sense of someone who subscribes to a view of freedom as the ability *to act otherwise* despite prior conditions of the universe remaining the same in all discernible ways. I also posit that the ability to act otherwise is a universal *capacity* of human beings, in the same way that language, for instance, is posited as a universal human capacity in generative linguistics[9]. As with any capacity, the *use* of the capacity to act otherwise may be stifled if certain external or (in the case of free human choice, internal) conditions are not met. Much the way a plant, for instance, has the capacity to blossom in the presence of water, warmth, and oxygen, certain external and internal conditions may be necessary to permit an individual to make full, or even partial, use of their capacity to do otherwise, i.e., to use their freedom. While not a lot is understood about the property or capacity 'to act otherwise', there is scientific work investigating the nature of voluntary action in humans that successfully isolates something I think we can describe as roughly representing this property or capacity. MIT neuroscientists Emilio Bizzi and Robert Ajemian in a paper entitled '*A Hard Scientific Quest: Understanding Voluntary Movement*s' offer a review of what is known about the neural processes underlying ordinary human actions (drinking coffee, lifting your arm, and so on). They ask in the opening section of their paper, how we can translate ''something as evanescent as a wish to move into muscle contractions'' further referring to this process as ''awfully complicated'' and ''only partially understood'' (p, 83).[10] The partial understanding that these scientists have achieved in explaining voluntary movement includes, amongst other things, understanding how the Central Nervous System – namely, the array of nerve tissues that control activities in the brain and spinal cord – has found simple computational solutions to controlling millions of muscle fibres simultaneously activated during the simplest voluntary actions.[11] How the human brain manages to coordinate the millions of muscle fibres involved in an action - in particular, the authors note, in those cases in which nearly all of one's muscles are being used simultaneously (the example they offer is an athlete playing soccer)- remains, they write, ''one of the fundamental questions in motor neuron science'' (p, 84).[12] To add to the complexity of muscle coordination, the authors highlight how little time one really has to pre-plan complicated actions, like shooting for a goal when playing soccer, concluding that there is plainly ''neither the time nor the inclination to explicitly formulate the command signals to control the millions and millions of muscle fibers'' in the body. They reason that the brain must therefore rely on an ''effective simplifying strategy'' (p, 84), and illustrate how progress has been made in discovering what that strategy is.[13] Bizzi & Ajemian also add, however, that it is not understood how the spinal cord initiates a movement in the first place. The authors write that ''no answers have yet been found to explain how the cortical areas of the frontal lobe construct the spatiotemporal patterns of neural activity necessary to activate the spinal cord, enabling it

[9] This is a basic concept in Chomsky's work on language. All it means is the presence of some innate ability that can be used by human beings as part of their cognitive apparatus. In the case of language, for instance, Chomsky posits that language is a computational system of human brains that has various uses (e.g., reasoning, communicating). This idea is taken for granted in Chomsky's work, but see, for instance, Ch 4 of *New Horizons in the Study of Language and Mind,* for an indication of how this concept underlies his work.

[10] Bizzi, Emilio, & Ajemian, Robert (2015), 'A Hard Scientific Quest: Understanding Voluntary Movements', *Daedalus*, Vol 144, No 1, pp, 83-95.

[11] Bizzi, Emilio, & Ajemian, Robert (2015), 'A Hard Scientific Quest: Understanding Voluntary Movements', *Daedalus*, Vol 144, No 1, p 93.

[12] *Ibid*.,

[13] *Ibid*.,

to execute a specific movement'' (p, 93).[14] According to these neuroscientists, we have, in other words, some understanding of how the spinal cord coordinates the muscle activations necessary for movement to occur, but no understanding of how the spinal cord initiates a movement in the first place. To illustrate the point, and to elaborate on the soccer example these neuroscientists have already used, we have some understanding of how a soccer player's nervous system permits coordinated - as opposed to, say, uncoordinated – muscle movement, but no understanding of how a soccer player decides when to move in the first place, or how, for example, a player decides to run to the left of an opponent instead of to the right when navigating the field while running. In summary, these neuroscientists argue that we have no understanding of how people act of their own volition.

I think this research suggests we can reasonably posit that something like the capacity to act otherwise is a property of human beings that has some empirical basis in the human nervous system (and perhaps the nervous system of other animals), though the capacity is not scientifically understood. Of course, the capacity to act otherwise is not entailed by Bizzi & Ajemian's research: the existence of a capacity in the Central Nervous System to execute particular movements in a way that is not understood, does not imply that these movements could have been otherwise in any immediate or obvious sense. On the other hand, if we couple this scientific result with the fact that we have (what I think we can reasonably take as) introspective evidence for our ability to act otherwise, there is some basis for defending the existence of freedom as an innate capacity of human beings. After all, for the overwhelming majority of the actions we take, we have the internal experience of believing that we could have acted differently and that we were the decider of our actions. This typically remains true even in cases where we feel our actions were heavily influenced by internal or external factors - sometimes illustrated in feelings like regret or remorse in cases where we are not proud of our actions, or in feelings of happiness or joy in cases where we are. Importantly, this intuition is not contradicted by any science. In fact, research supports that the property appears real from an empirical perspective, though it also eludes the neurosciences.

If we define freedom as a capacity of human beings in the sense I am suggesting, we can also distinguish two related but often conflated questions: the question of whether free will exists as an empirical matter of fact, and the question of whether an individual was really free under 'such and such' circumstances in acting as they did. In other words, I propose that we distinguish freedom, defined as something like the ability or capacity to act otherwise, from our *use* of that freedom, which under certain external or internal circumstances may be impaired. It would then be true under different circumstances that one is more or less 'free', or perhaps entirely 'unfree', without this implying that the ability to act otherwise is absent or fails to exist. To see why I suggest drawing this distinction, consider, for instance, the definition of freedom used in Libet's work to examine free will. Libet outlines a definition of freedom that he describes as ''in accord with common views'' (p, 47)[15] that must meet two conditions. First, there should be ''no external control or cues to affect the occurrence or emergence of the voluntary act under study'' (p, 47); second, the subject must feel in control of their action when it is being done, feel that they want to do the action, and feel that the action is of their own initiative.[16] While these conditions, and I think others, should be considered in deciding how free one is under different circumstances to act in accordance with their

[14] *Ibid*.,

[15] See Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, pp, 47-57.

[16] See Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, p 47.

will, they strike me as overly restrictive in trying to establish whether free will exists (at least if one accepts my definition of freedom). Suppose for instance, an individual puts a gun to someone's head and tells the person to rob a bank. For all practical purposes this person is not free under such circumstances, because the use of their ability to act otherwise, i.e., to not rob the bank, is *impaired*. Nevertheless, I think it remains theoretically possible that the person chooses not to rob the bank, however unwise it might be to choose that action under the circumstances, or however impossible it might be to choose to act differently under the circumstance. One could very easily imagine, for instance, being so overcome with fear in that situation that one does not even think to disobey the instructions to rob the bank. In that case, one's fear is so great that their ability to act otherwise is impaired, without this implying that free will does not exist. In other words, the *relevant question* in such circumstances is not whether one's ability to act otherwise exists, but how much the external and internal conditions of the situation permit one to make use of that capacity.

According to Libet's definition of freedom, however, the conditions listed are one's that have to be met for free will to exist. In a paper entitled '*Do We have Free Will*?', Libet set out to test whether we have free will in the robust sense that the conditions he enumerates at the beginning of his paper require. In these experiments, subjects are asked to flick their wrist whenever they feel like doing so, and the brain's 'readiness potential', defined as ''electrical indication of certain brain activities'' (p, 49)[17] is measured. It was consistently found across several trials that readiness potential reliably preceded any conscious awareness of the will to act by about 350-400 milliseconds, and preceded an action by about 200 milliseconds.[18] In other words, the brain shows signs of preparing for action before one can articulate that they intend to act and before they do act. As Libet puts it, ''the initiation of the freely voluntary act appears to begin in the brain unconsciously, well before the person consciously knows he wants to act!'' (p, 51).[19] Libet appears to conclude from these experimental results that the second condition on freedom, that is, the condition which says one must be in control of their action, feel they want to do the action, or initiate the action themselves, has not been met. He suggests that for that condition to be met, one would need the brain's readiness potential to emerge at approximately the same time as the conscious intention to act, claiming that ''if conscious will were to follow the onset of [reading potential], that would have a fundamental impact on how we could view free will'' (p, 49).[20] Libet himself did not take these experimental results to show that free will does not exist. Rather, he took them to imply a more restricted definition of freedom than he began testing with: we can no longer think of freedom as the voluntary formation of an intention to act, given, he appears to believe, that an intention to act is *involuntary* if it is *unconscious*. Instead, he argues that we should view freedom as the ability to veto the occurrence of certain acts. According to his experimental trials, the conscious will to act appears about 150 milliseconds before muscle activation, which, he reasons, would leave enough time for the conscious will to veto an act. While Libet's experiments represent interesting results for the neurosciences - in particular, for thinking about the conscious-unconscious divide - my own view is that Libet winds up with an unnecessarily restrictive definition of free will, as merely the ability to *veto* an act rather than *choose* one, in part because the definition of free will he begins with is also too restrictive.

Before discussing this further, however, let us look at Libet's interpretation of his

[17] See Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, pp, 47-57.
[18] See Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, pp, 47-57.
[19] *ibid*.,
[20] *ibid*.,

experimental results, and how his interpretation of the results implies that the second condition he places on freedom, is not met. Critical to Libet's interpretation of his results is the belief that the formation of an unconscious intention to act implies that the intention to act was not voluntary or within control of the agent. The reasoning appears to be something like this: suppose we were able to slow down time such that we had enough time to ask an individual during the 350-400 milliseconds in which their brain showed readiness potential to act, whether they in fact intended to act. Libet seems to think that this person's response would be 'no', despite the fact that 400 milliseconds later this very same individual would act. The negative affirmation of an intention to act would then reasonably imply that it was beyond the person's control to form the intention to act. After all, if we assume that the brain's readiness potential implies that the person has formed an intention to act, and there is a particular time period in which the individual would report no intention to act, one may reasonably conclude that the person did not *choose* to form an intention to act. That is to say, their brain processes were operating 'against their will'.

There are, however, other ways to interpret these results. It could just be, for instance, that it takes approximately 350-400 milliseconds for the formation of an intention to reach consciousness. In other words, if we slowed down time and asked the person whether they intended to act, we might learn nothing of interest, and witness only a slow-motion image of someone's brain working away at a subconscious level making decisions, and a person then articulating these decisions 400 milliseconds after they began to take form. After all, it is also being assumed in the first interpretation of these results, that 'readiness potential' is enough to establish that a decision to act one way or another has already been formed, though we might read it as only establishing that the formation of a decision to act is underway. Given that readiness potential is only a measure of electrical brain activity in parts of the brain to do with action, and given that the subjects in this experiment have already been primed to act within some reasonable time frame (assuming they do not want to leave the experimenter waiting all day to gather results from the test trials), it is not obvious that brain activity implies that a decision to act a certain way has already been made within these 400 milliseconds. In that instance, were we able to slow down time to ask the agent if they were going to act, one might plausibly discover that the agent simply has no response, that it takes approximately 400 milliseconds to respond. Suppose, however, that readiness potential does imply the formation of a decision to act. Even still, it presumably takes some time for the brain to articulate thoughts to consciousness. In other words, for the same reason Bizzi & Ajemian suggest there is 'neither the time nor the inclination' for the nervous system to explicitly formulate conscious commands about muscle coordination, it may also be that the brain decides how to act and only later, if at all, has the inclination to communicate this act to awareness. In fact, many actions that we intuitively take for granted as voluntary, such as picking up a glass of water to drink, or moving our fingers to type on a screen, are actions that we take without necessarily articulating the decision to act to ourselves before acting. This might suggest that an intention to act need not always coincide with a conscious articulation of an intention to act.

If we interpret Libet's results in this second way, should we conclude that the person's intention to act was not voluntary? I don't see any particular reason to do so. On this interpretation, one's intention or decision to act would merely be operating at a lower level of awareness than consciousness, as opposed to operating in a way that was in conflict with the person's subjective reportable intentions. Suppose, however, that we interpret Libet's results as showing something greater than that, as showing, for instance, that the formation of intentions to act are to some extent beyond our control, because they happen at an unconscious level, or because they happen

in a way that renders it theoretically possible that, at some early stage of brain processing, our subjective reporting is in conflict with our brain activity. If we define freedom as nothing more than the capacity to act otherwise, and as a capacity *for* action, as opposed to a capacity for formulating intentions, then Libet's experiments do not threaten the existence of freedom. In fact, many aspects of the workings of our psyche are beyond our control in some sense. We cannot prevent ourselves, for example, from having certain thoughts or feelings pop into our heads, nor can we prevent our brain form thinking 24/7 - though we can train ourselves to quiet our thoughts - without us taking such facts to threaten our ability to choose otherwise. In other words, even if one believes the formations of intentions are beyond our control in certain senses, the crucial point for freedom is that we have the ability to choose whether we act on those intentions (or in more neutral terms, whether we act on those thoughts).[21] Interestingly, Libet himself seems to agree with this claim in large part, documenting on multiple occasions throughout his paper that freedom in this sense does exist. He notes for instance that ''the subjects in our experiments at times reported that a conscious wish or urge to act appeared but that they supressed or vetoed that'' (p 52).[22] Nevertheless, given the definition of freedom Libet begun with, he ends up refining his definition of freedom to the more restrictive notion of the ability to veto choices, where he conceives of this ability to veto as a ''control function'' (p, 53).[23] This is to avoid the possibility that the vetoing of an action also has its roots in unconscious thought processes that are involuntary. While I think Libet's conception of freedom as the ability to veto an action is certainly more preferable than assuming determinism threatens the ability to control one's actions at all, I do think his conception of freedom is unnecessarily restrictive, in part at least, because of how he defined freedom at the outset of his experiment.

3 Does Determinism Threaten Free Will?

If we grant the conception of freedom delineated above for the reasons I have suggested, is it really true that the sciences, and in particular, deterministic systems, generate a compatibilist incompatibalist debate? To answer this question I would like to shift our attention away from the metaphysical thesis of determinism or any other kind of metaphysical picture we may attach to scientific theories of the world, and focus instead on delineating a conception of 'science' that puts aside metaphysical concerns about how the world is, and looks instead at what is strictly inferable from the results the sciences have achieved. While it is common to conceive of the sciences as inherently tied to certain metaphysical assumptions about the world, it is not unheard of to reject this kind of approach. Noam Chomsky, for instance, conceives of the sciences as ''whatever is achieved in pursuing naturalistic inquiry'' (p, 81)[24], further stating that he understands naturalism – namely, science, ''without metaphysical connotations'' (p, 81).[25] A naturalistic approach to the mind, he writes, ''investigates mental aspects of the world as we do any others, seeking to construct intelligible explanatory theories, with the hope of eventual integration with the [core] natural sciences'' (p, 76).[26] It is worth noting how narrow Chomsky's conception of science really is: he is advocating a picture of the sciences as no more (or less) than the explanatory theories the sciences produce.

[21] What an 'intention' is, remains a notoriously difficult problem in philosophy and cognitive science. This complicates Libet's experiments in certain ways, but I leave this question aside here.
[22] See Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, pp, 47-57.
[23] See Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, pp, 47-57.
[24] Chomsky, Noam (2000), *New Horizons in the Study of Language and Mind,* Cambridge University Press.
[25] *ibid.*,
[26] *ibid.,*

According to such a view, we therefore take in our stride what the sciences explain and do not explain about the natural world, and whether the sciences, so understood, imply a unified metaphysical worldview of the kind that is generally assumed to have emerged with Newton. That is, a metaphysical picture of the universe as determined, mechanistic, and physical.

How, you might ask, does this conception of science differ from the standard conception underlying the determinist-free will debate? The traditional metaphysical vision underlying this debate is what is sometimes called *necessitarianism*[27]. This view claims that the laws of physics somehow 'govern' the universe in a way that future states of the universe are caused or necessitated by prior states of the universe.[28] Perhaps the most well-known articulation of this view in the history of determinism is LaPlace, who argues that we ought to ''regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow'' (p 12).[29] Importantly, Laplace further claims that we can think of the past as entirely determining the future, on such a worldview, provided the intelligence underlying the equations of physics are all-knowing. He writes that for an ''intelligence which could comprehend all the forces by which nature is animated and the respective situations of the beings who compose it […] nothing would be uncertain, and the future as the past, would be present to its eyes'' (p, 12).[30]

Not one but two core assumptions underlie this determinist or necessitarian thesis. First, we have to assume that the laws of physics along with previously determined states of the universe, necessitate, determine, or cause future states of the universe. Second, we have to assume that the laws of physics, and future and past states of the universe, imply an understanding of everything there is to know about the world, thus comprehending, one assumes, the nature of human action. Some compatibilists have rejected LaPlace's conception of scientific laws as including an element of necessity. Bernard Berofsky, for instance, in a book entitled *Nature's Challenge to Free Will* claims that ''freedom is compatible with determinism because the only reason to believe otherwise is based on the false metaphysics of *necessitarianism*'' (p, 3)[31]. On the other hand, rarely do we witness the rejection of the second assumption embedded in LaPlace's discussion of determinism, namely, that there exists a complete explanatory description of the world.

If we reject both of these assumptions, I think we can move beyond the free will-determinist debate by reasoning as follows. While it may be true that a descriptive explanation of the motions of objects is a complete explanation of the motions of objects, it is not obvious that a complete description of the motions of human beings is a complete explanation of human beings, including, say, their ability to act otherwise. This might appear to generate some sort of logical contradiction: if deterministic systems can in theory offer a description of the whole state of the world at time Tn, does that not imply that human actions are also determined to be a certain way at time Tn? To generate this contradiction we need to believe in more than just the laws of physics and their ability to describe in totality the state of the universe at any particular time, we also need to believe that these laws, in virtue of their explanatory power, are somehow unbounded in scope. In other words, we need to

[27] See, e.g., Berofsky, Bernard (2012), *Nature's Challenge to Free Will* (Oxford, 2012; onlinedn, Oxford Academic).
[28] This view is also rejected by some who discuss the issue outside of the free will determinism debate. See, for instance, Albert, Z, David, (2015), *After Physics,* Harvard University Press.
[29] See, Laplace, Pierre-Simon, *A Philosophical Essay on Probabilities*, Translated from The Sixth French Edition, by Frederick Wilson Truscott & Frederick Lincoln Emory, *First ed*, John Wiley & Sons: London, p, 12.
[30] *ibid.,*
[31] Berofsky, Bernard (2012), *Nature's Challenge to Free Will* (Oxford, 2012; online edn, Oxford Academic).

assume that anything that is subsumed under physical laws (which we assume everything is) are also *explained* by these laws. We can, however, question that assumption. While the motions of humans is surely subsumed in some sense by the laws of physics -assuming, that is, that human beings are biological creatures - does this really mean that the laws explain human action?

To return to Bizzi & Ajemian's soccer example, does describing where on the field a soccer player is at various intervals of time, explain how the soccer player gets to these different positions on the field at various intervals of time? If the soccer player were a rock, such that the player lacked a human nervous system and brain that appears to suggest it could have acted otherwise, then I think the answer to this question is 'yes'. A total description of the universe would explain where a rock would land at various time intervals on a field in a way that explained the motions of the rock. The reason is that since the advent of Newtonian physics, we have scientific laws that allow us to calculate the motions of things like rocks, in a way that predicts the behaviour of these objects (though today, with the advent of quantum mechanics more precise probability distributions may be needed for accurate prediction). This explanation is successful, however, because we only need to know the rock's mass, position, and the forces acting on it, to explain its behaviour. I do not believe, however, that knowing the mass, position and the forces acting on a soccer player explains how a soccer player chooses to act under certain conditions on the field. If one of the relevant properties to explain in the case of human action is the ability to act otherwise, then the requisite property simply does not fall within the purview of current physics. If tomorrow, for instance, someone was to offer us a full physical description of the world along with a unified science that connected everything that is known in physics, chemistry, and biology, with everything that is known in motor neuron science, the nature of human action would still elude our understanding. The reason for this is because the motor neuron sciences do not understand how human action works. To believe, in other words, that we can explain human action in terms of the mass, position, and forces acting on human beings, is *already to assume* that there is nothing the laws of physics cannot explain. Indeed, this assumption is generally baked into the conception of science or metaphysical vision of reality that underlies determinism.

On the other hand, if we view the sciences as nothing more (or less) than the explanatory theories that the sciences produce, we turn the question of whether the sciences explain human action into an empirical question in its own right. When we do this, what we find I think, is that the sciences do not explain the nature of voluntary action or free will. It is worth noting that others have plainly recognised that physics does not explain free will. Libet himself, for instance, who is often read as concluding that there is no such thing as freedom, writes that ''non-determinism, the view that conscious-will may, at times, exert effects not in accord with known physical laws is of course also a non-proven speculative belief'' (p, 56), further claiming, regarding the evidence for determinism precluding free will, ''that such evidence is not available, nor do determinists propose even a potential experimental design to test the theory'' (p, 56).[32]

One could summarise my position as reiterating that it is indeed open to us not to indulge - in what Libet terms- this 'speculative belief'. Instead we can remain agnostic on the question, until and *if* evidence to the contrary emerges, on whether the sciences has anything of interest to say about freedom. This perspective on the relationship between physics and free will is generally lacking in the compatibilist incompatibilist debates on freedom: Even the

[32] See Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, pp, 47-57.

most sophisticated compatibilist positions do not *ask* whether the sciences threaten free will, they *assume* that they do. If we make this move, we then proceed to focus our attention on reconciling free will with determinism according to our preferred metaphysical picture of reality. On my view, however, there is an open path for libertarians to believe in free will that is not secured through quantum uncertainty. Rather, it is secured through narrowing the explanatory scope of the sciences – and therefore the explanatory scope of determinism or (what we might call) probabilism. This requires accepting instead that free will does not automatically fall within the explanatory purview of the sciences, leaving it an open question whether a complete and true physics will explain this property.

REFERENCES

[1] Albert, Z, David, (2015), *After Physics,* Harvard University Press.

Berofsky, Bernard (2012), *Nature's Challenge to Free Will* (Oxford, 2012; online edn, Oxford Academic, 24 May 2012), https://doi.org/10.1093/acprof:oso/9780199640010.001.0001, accessed Dec 10th 2024.

Bizzi, Emilio, & Ajemian, Robert (2015), 'A Hard Scientific Quest: Understanding Voluntary Movements', *Daedalus*, Vol 144, No 1, pp, 83-95.

Chomsky, Noam (2000), *New Horizons in the Study of Language and Mind,* Cambridge University Press.

Chomsky, Noam, (2016), 'What kinds of Creatures Are We?' *Columbia University Press.*

Chomsky, Noam (2006) *Language and Mind* Chomsky, Noam, (3rd ed.). Cambridge University Press.

Chomsky, Noam (2009) *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*: Vol 3rd ed. Cambridge University Press.

''Chomsky, 'Free Will and indeterminism'' uploaded by *Chomsky's Philosophy,* posted March 14th 2017 https://www.youtube.com/watch?v=92UjJ8p7cJs

''Noam Chomsky, Free Will I'' uploaded by *Chomsky's Philosophy,* posted Sept 29th 2014. https://www.youtube.com/watch?v=J3fhKRJNNTA,

Explaining Consciousness: The Hard Problem (1997)., edited by Jonathan Shear. MIT Press.

Holbach, Baron, 1770, "The Illusion of Free Will", from *The System of Nature* (*Système de la nature*). Translated by H.D. Robinson. Reprinted in *The Experience of Philosophy*, Daniel Kolak and Raymond Martin (eds.), Oxford: Oxford University Press, 2002, 5th edition: 176–181.

Hume, David (1975) *An Enquiry Concerning Human Understanding* P.H. Nidditch (ed.), Oxford: Clarendon Press.

Laplace, Pierre-Simon, *A Philosophical Essay on Probabilities*, Translated from The Sixth French Edition, by Frederick Wilson Truscott & Frederick Lincoln Emory, *First ed*, John Wiley & Sons: London

Libet, Benjamin, (1999), 'Do We Have Free Will?', *Journal of Consciousness Studies* 6, No. 8–9, pp, 47-57.

Punishing the Innocent?

The Normative Impotence of Hard Determinism

Thomas Nys[1] & Gijs van Donselaar[2]

Abstract. Hard determinists, like Galen Strawson, believe that their theory has normative repercussions, that is, they maintain that the truth of determinism commits us to a certain normative position, They say, for instance, that legal (criminal) punishment on the basis of desert – the Retributivist Perspective on punishment – is unjustified. We should instead adhere to a different principle for administrating punishments and rewards. Saul Smilansky, however, has argued that the adoption of the hard determinist's stance on punishment would lead to 'funishment' and, consequently, to the collapse of society. In this contribution, we will argue that both Strawson and Smilansky are wrong. There is no moral force to hard determinism, so nothing, let alone 'funishment', follows from it.

1 INTRODUCTION

One of the central and 'classical' principles of criminal law – still believed by many[i] - is that punishment is due to the guilty, not the innocent. The guilty *deserve* punishment, whereas the innocent do not. Some philosophers, however, have questioned this principle arguing that there are no grounds for such moral desert, thereby invalidating this particular foundation for legal punishment. Galen Strawson, for instance, in a famous and influential paper[ii], argues that punishment is never justified, at least to the extent that is not deserved. No-one deserves punishment, so all punishment on the basis of desert is unjust. And some other 'hard determinists'[iii], like Derk Pereboom, have therefore concluded that criminal law, or the normative bedrock on which it rests, should be seriously revised or reconsidered. This view, however, has not remained unchallenged. Saul Smilansky, for instance, argues that the adoption of this framework would lead to a 'practical reductio' and therefore to a defeat of the hard determinist's position on punishment.[iv]

In this paper, we want to disagree with both Strawson and Smilansky. Galen Strawson (and other, like-minded theorists) is incorrect in concluding that hard determinism would prove the injustice of punishing the innocent. In fact, the very idea that his argument should have *any* practical consequences is self-defeating. Smilansky, on the other hand, is wrong to suppose that hard determinism requires punishment to be replaced by 'funishment' and that its normative consequences would therefore lead to a collapse of society. Both authors are wrong for the same reason: there is no moral force to hard determinism. No moral 'ought' follows from it.

2 STRAWSON AND SMILANSKY

Let us first give a brief but more detailed account of Strawson's and Smilansky's positions. Galen Strawson gives a seminal exposé of so-called 'hard determinism': the position that, because of the truth of determinism, we lack the free will ultimately needed to support moral responsibility. His argument starts from 'psychological determinism'. Since our actions are determined by who we are, that is, by our desires, purposes, and beliefs, we can only be responsible for our actions if we are also responsible for who we are. But if it were true that we somehow brought about who we are this must have been on the basis of who we were before,

[1] Associate Professor of Ethics, Department of Philosophy, University of Amsterdam, t.r.v.nys@uva.nl.
[2] Assistant Professor (retired), Department of Philosophy, University of Amsterdam.

and therefore we enter into an infinite regress, the ultimate bugbear of philosophical argumentation.[v] So, a simple version of what Strawson calls the Basic Argument runs like this:

(1) Nothing can be *causa sui* – nothing can be the cause of itself
(2) In order to be truly morally responsible for one's actions one would have to be causa sui, at least in certain crucial mental aspects
(3) Therefore nothing can be truly morally responsible[vi]

Strawson concludes that we are therefore not ultimately responsible for our actions. As a consequence, it would be unjust to punish criminals for their crimes because they 'could not have acted otherwise'; they are not the ultimate cause of who they are (and how they decided to act). As he puts it:

"[T]he evident consequence of the Basic Argument is that there is a fundamental sense in which no punishment or reward is every ultimately just. It is exactly as just to punish or reward people for their actions as it is to punish or reward them for the (natural) colour of their hair or the (natural) shape of their faces."[vii]

Since we do not think it is justified to punish people for the color of their hair or the shape of their faces, we should extend this judgement to all human actions. People do not *deserve* punishment.

Saul Smilansky elaborates on this – what he calls the "moral force of hard determinism" – and argues that it implies a 'practical *reductio*'. We may detain criminals, he argues, just as we put people with contagious diseases in quarantine, but since they are innocent and do not deserve their loss of liberty, we should compensate them and turn jailhouses into fancy five-star resorts. Since it is, as Strawson surmised, unjust to punish the innocent we should only subject them to 'funishment'. But in that case the deterrent effect of incarcerations will go down because funishment is not nearly as bad as punishment. In fact, there could even be a win/win situation for criminals: if they get caught, they will be sent off to a five-star resort; and if not, they will reap the benefits of their crimes.[viii] Such a penal system will not only become unbearably expensive, but, in the end, society will collapse under rampant crime.[ix]

In what follows, we will argue no such collapse is to be foreseen if hard determinists, on a proper understanding of determinism, adopt a moral system that is compatible with it. This is where Smilansky is wrong. But next we will question the moral force of determinism itself by showing that it leads to a contradiction. And this is where hard determinists like Strawson Jr. and Pereboom are mistaken.

3 THE CATEGORY MISTAKE

It is true that people are innocent of a crime if they were not able to commit it. In some court cases an accused person's innocence is proven if it is established that she was not even near the crime scene. On December 31, 2008, for instance, *The New York Times* reported how a murder charge against a man, named Jason Jones, was dropped after his lawyers had established, with the help of information on Jones' New York metro card, that he was actually miles away from the crime scene at the time of the murder. So, Jones could not have done it; he did not have the opportunity to do it, and therefore he did not do it. He was innocent. We might even put it in the following cumbersome way: Jones was innocent because he 'could not have acted otherwise' than *not* committing the crime.

Hard determinists exploit this ordinary understanding of innocence (about which there is nothing 'metaphysical' except the assumption that people cannot be in two places at the same time) when they argue that people who *did* commit a crime are also innocent because, being determined, they 'could not have acted otherwise'.

The parallel is this:

If people are innocent in case

(a) they could not have acted otherwise than not committing the crime (not being near the crime scene); they were unable to do it

they are also innocent in case

(b) they could not have acted otherwise than committing the crime (as when they actually did and were determined to do so); they were unable not to do it.

This parallel, however, is flawed. The 'could not have's' are logically distinct between the cases. The 'could not have' in the first case is not unqualified; it is conditional on the accused person not being near the crime scene. In case she had been at the crime scene at the time of the crime she could either have committed the crime, or not. Perhaps she would not have committed the crime anyway, but she *could* have committed it. On the ordinary understanding of innocence, we are not implying that it is impossible for her to commit crimes or that she is incapable of committing them. She is innocent merely in this particular case, given the evidence.

For the hard determinist's 'could not have' there is no such condition or qualification. People could not have acted otherwise, whether or not they actually committed a crime. Moral innocence requires no specific alibi. For hard determinists the 'fact' that people could not have acted otherwise is not refutable by empirical evidence. It is metaphysical. There is no possible world in which the same natural laws would apply and criminals would be guilty. But from this it rather follows that the verdicts of 'guilty' or 'innocent' are not meaningfully applied to people. The suggested parallel amounts to a so-called 'category mistake'.

Consider another parallel that we hope is useful. When we say of a rock that it is not alive, we cannot be taken to say that the rock is dead. If that were so, one could meaningfully ask when the rock died, or how long it lived. We rather try to convey that rocks cannot be alive *or* dead; they are objects to which the concepts of life and death do not apply. Neither can rocks be happy or depressed, or brave or cowardly, or right or wrong.

Likewise, when a determinist says that criminals are not guilty for their actions, it cannot be implied that these criminals are *innocent*. Instead the determinist should saw that human beings, lacking free will and real choice, are not capable of being guilty or innocent. Indeed, just as it is a mistake to say of rocks, or rivers, or of animals, that they are guilty or innocent. They are all, so to speak, '*beyond* good and evil' (or 'before', if you like). The hard determinist is right in saying that it is not justified to punish them, but not punishing them would also be unjust, or rather, non-just (for people who claim that rocks are alive are as much mistaken as those who label them dead).

What then of the moral force of determinism, the view that it would be unjust to punish criminals? If, according to determinism properly understood, we incarcerate criminals and make their lives unattractive in the traditional way, we are *not* punishing the innocent because criminals, like all others, and like rocks, rivers or animals, are guilty *nor* innocent.

Let us put the same point differently. Criminals, the hard determinist says, do not deserve to be punished because they are not responsible for their actions. But what *do* they deserve? What do they deserve because of something that is similar to the color of their hair, or the shape of their faces? Nothing. If all choice and behavior is determined, then nobody deserves anything, ever, just like rocks, rivers and animals do not deserve anything. So, the determinist cannot claim that punishment for criminals is either deserved or not deserved. The impossibility of moral responsibility implies the impossibility of moral desert *tout court*.

Perhaps this is even clearer if we consider the reverse of blame and punishment: praiseworthiness and reward. Consider a medical doctor who goes out of her way to care for her patients. She is meticulous in making diagnoses

and does her utmost to prescribe therapies tailored to each patient's condition. She regularly visits them to monitor their progress, offering mental support to both her patients and their families. Moreover, she has established a fund to provide poor orphans with access to medical care. Yet, despite her dedication, some people in the community – resentful because she is an 'outsider' rather than a local – spread malicious gossip about her. This baseless slander tarnishes her reputation and ultimately destroys her career in that community.

Normally, one would say that her sorry fate is undeserved, and that she is wronged by those vicious gossipers for that reason. But a hard determinist cannot say that. According to the hard determinist she could not have acted otherwise than in her admirable ways, and therefore there is nothing she deserves, or does not deserve, simply because of her conduct. She may not deserve her treatment by the gossipers[x], but for the hard determinist there cannot be something else that she does deserve (e.g., praise).

Let us return to the notion of 'punishment' itself. We do not punish rocks, rivers, or even very dangerous animals. We may do all kinds of things to contain or control or destroy them if we have reason to, but these things we do are not punishments, as these treatments are not deserved or undeserved. So, if we leave the retributive, or deontic, element out of our justification for the way we treat criminals – the idea that we give them something they deserve – it is wrong to call that treatment a 'punishment'. The concept of retributive punishment, like those of responsibility and desert, like those of guilt and innocence, or praiseworthiness, does not apply to beings that are determined in their actions.

4 FROM HARD DETERMINISM TO CONSEQUENTIALISM

So, *if* hard determinism has any moral force, it appears that it can only be compatible with a moral theory that itself has no concern for punishment, at least not of the retributive kind that speaks in terms of moral responsibility, blame and desert, and guilt or innocence. One articulate moral theory that does so is, of course, consequentialism. A rigidly consequentialist regime, for instance, may call the populace to Main Square announcing that *this* is what will happen to any person who breaks the rules and then proceed to execute a randomly selected citizen in a particularly brutal way, for all to witness. If that citizen laments that she has never broken any rule and that she is treated unjustly, this regime will reply that the question of desert does not enter the moral equation in the first place and that her execution is justified as long as the spectacle is sufficiently appalling and serves as a deterrent for others.[xi]

For consequentialists, therefore, it also does not matter whether or not the deprivation that we visit on people we put in detention because they behave in socially harmful ways, is deserved. Consequentialists just have to optimize the balance between the dismay of the deprivation and the benefit of society as a whole. And if we do so there is nothing unjustified about what we do (in fact, that is what we *should* do, our moral 'ought'). And therefore no-one deserves to be compensated for being detained or otherwise ill-treated. There is no need for funishment, and a consequentialist society will therefore not collapse under the wight of having to compensate criminals.

Smilansky explicitly considers the possibility of a consequentialist response to his practical *reductio* of hard determinism, but he rejects it:

> "[A] hard determinist arguing in this way would betray the moral force of hard determinism. Utilitarianism has, of course, been available for a long time, but people worried about free will and moral responsibility are concerned with more than utility. This concern has been the basis for taking the moral high ground as against a solely utilitarian justification of blame, guilt, and harsh punishment, rightly saying that making people

> suffer guilt or punishment just because doing so would be socially useful is morally unacceptable. We cannot use the morally innocent in such ways even if it furthers social interests. Blame and punishment must also be just, and not only socially efficient. And in order for them to be just, they must follow upon the choices and actions of moral agents, who through their free actions have made themselves liable to blame and punishment. That is why the punishment of the innocent is the paradigm of injustice (yet it is of no principled concern to a utilitarian). But for hard determinists, everyone is morally innocent! Hard determinism as a moral position thus holds that no one deserves to be made to suffer, or to be made worse off than another, and hence that it would be unjust to do so. Hard determinism cannot turn to consequentialism for assistance in overcoming the *reductio*, for it would thereby completely betray itself as a distinct ethical position."[xii]

The crucial sentence is in the middle of this long quote: "we cannot use the morally innocent in such ways [punishing them] even if it furthers social interests".[xiii] This observation would be correct if punishment were applied to the morally innocent, but as we argued, the set of the morally innocent does not include all of us (as Smilansky invites us to believe). Instead, according to hard determinism properly understood that set is *empty*, as is the set of the morally guilty (we are all 'no-nocent', we would say; the terms of innocence and guilt do not apply to us). So, what the 'distinct ethical position' of hard determinism would be, if it not consequentialism, remains a question.[xiv]

Consequentialism and hard determinism are definitely compatible. Consequentialists, for example, could be hard determinists, but they need not be. In fact, the question of free will need not concern them (as they are all about the consequences).[xv] Similarly, hard determinists might embrace consequentialism (for example, to navigate out of the practical reductio posed by Smilansky), but they need not be.[xvi] In fact, what we will argue in the remainder of this paper is that hard determinists are not committed to any normative position *whatsoever*.

Hard determinists can endorse consequentialism, but nothing in the truth of hard determinism *commits* them to that normative framework. Consequentialism is just an ad-on; a convenient tool to somehow rescue – albeit in a modified version – the practice of praise and blame, punishment and reward. After hearing the hard determinist's exposé, a surprised or shocked member of the audience could say "What? So that would mean that we should no longer punish our criminals?!", and the hard determinist would reply, "No, not necessarily, because there are *other* grounds for 'punishing' people."[xvii] But these are indeed entirely different grounds. What we ought to do is independent of the truth of hard determinism; no 'ought' follows from it. In fact, no 'ought' can follow from it, for to claim that it does, would be self-defeating.

5 HARD DETERMINISM'S NORMATIVE IMPOTENCE

Suppose a person fails to act on a principle that follows from the hard determinist's distinct moral position, whatever that is (e.g., consequentialism). Then, according to the hard determinist she is not guilty of, or responsible for, or deserving of, anything, because she could not have acted otherwise. But if he could not have acted otherwise, then how can it be that he *ought* to have acted otherwise. After all: ought implies can. Therefore, the moral *force* of hard determinism, whatever its distinct moral position, destroys itself.

Of course, consequentialism, in and of itself, may provide grounds for non-retributive punishment. But hard determinism does not imply consequentialism; the idea that one *ought* to maximize good outcomes. If ought implies can, then all oughts should be suspended (we will return to this below, following Nietzsche), or – alternatively – everything can remain just the way it is. As Peter Strawson already noted: "If to all offenders, then to all mankind."[xviii] If criminals should not be blamed for their awful acts, then

surely we should not be blamed for blaming them. And this goes for all systems of blame or punishment (however 'just' or 'unjust' we might consider them). There is no *other* ground for condemning those who either reject consequentialism or fail to act on its principles other than the truth of consequentialism itself.

Hard determinists are not totally unaware of this problem. Some, for example, call the charge about the injustice of the retributive punishment principle their 'last accusation'[xix], because they realize that once the truth of hard determinism is revealed it undercuts the foundation from which to raise such accusations. Be that as it may, even this last accusation is one accusation too many. Hard determinists are simply not in a position to offer specific moral recommendations.

We can unpack this further. What about other than moral oughts? Imagine a student who has an important exam the next morning and cares deeply about passing. He knows he won't perform at his best if he gets drunk the night before. Yet, despite this, he takes a few more shots of tequila. The next day, he shows up to the exam in a deplorable state, having forgotten nearly everything he worked so hard to study, and fails miserably. Determinism holds that the student, in choosing to drink excessively, could not have acted otherwise; therefore, it cannot be said that he should have acted otherwise. So, we have no reason – in fact, in would be un-just – to accuse the student of being imprudent, or silly, or irrational, or whatever; as these qualities or predicates do not apply to rocks, rivers and animals either. Surely, we can say that it would have been better, and better *for him*, had he refrained from drinking too much, but given the way he was (his psychological make-up, his character), and given the circumstances, he is not to blame for anything.

Or consider a person who believes that she is a human being and that human beings are mortal, yet also believes that she has life eternal (the Ivan Ilyich scenario). Ought she to believe otherwise? No. For she *cannot* believe otherwise if the *does not* believe otherwise, so we shouldn't call her names that we would not give to rocks, rivers and animals, like 'incoherent' or 'inconsistent' or even 'stupid'. So, determinism cannot find sufficient footing to deliver any normative recommendations at all. Not only when it pertains to morality or law, but also in the realm of prudence and the rules of 'sound thinking'.

If this is true, then hard determinists find themselves in a rather intriguing and embarrassing position. Consider again Galen Strawson's paper entitled 'The Impossibility of Moral Responsibility'. In it a 'basic argument' is developed that argues from the truth of psychological determinism to the practical conclusion that it would be unjust to punish criminals (as it is unjust to punish them for the color of their hair).

So, the determinist author, believing that she has given an *argument*, and a basic one at that, must (1) think that the reader ought to accept the conclusions from that argument, even if that conclusion is uneasy to live with. But if the reader does not accept that conclusion, because she simply is who she is, she is also not to be blamed for not 'buying the argument' and the determinist must now (2) be thinking that it is not the case that the reader ought to accept the conclusion and act upon its moral force, because she cannot believe otherwise. If in response the determinist rejects (2) she is no longer a determinist. But if in response she rejects (1) that the reader ought to accept the conclusion from a sound argument, and act upon it, then we may ask: why spill so much ink in giving the basic argument in the first place?

The determinist may reply to this that from common experience she has learned that many people happen to have a psychological tendency to be responsive to a syllogistically arranged series of linguistic utterances, and that this makes them act in certain ways. Therefore, in exposing such an arrangement, she hopes to influence and steer as many, who are thus psychologically inclined, as he can.[xx] The basic

argument is addressed, so to speak, to those who are 'manageable' by syllogisms.

But that answer is odd. Most people comply with an uneasy conclusion drawn from a sound argument precisely because they believe they *ought* to do so; experiencing its normative force, not because they are somehow 'prodded' to the desired end. Galen Strawson mentions our Biblical obsession with heaven and hell as a reason for believing in basic desert, the belief he wants to invalidate by using the basic argument, but if that obsession is not enough to warrant the belief, then no other obsession – our psychological proneness to respond to reason; many humans being 'syllogistically obsessed' – could take its place.

The hard determinist who rejects (1) therefore admits that there is no moral force, or any force, emerging from the basic argument. In that case, the basic argument would be exploiting the current beliefs (the psychological set-up) of those to whom it is addressed. Can it then be called an argument at all, rather than a kind of manipulative mental 'massage' for the author's own purposes: taking the most efficient means to goad the masses to a better situation? But then, again, consequentialism is the solid, ground-level normative theory from which to argue, and hard determinist 'talk' is just a means in the consequentialist toolbox to achieve its goals.

Friedrich Nietzsche, who is embraced as an ally by Galen Strawson in his paper, at times understood the upshot of his own position very well. If 'guilt' and 'blame' are inventions of a slave morality, then we might – in a first move – reject these labels as fraudulent make-belief (the fierce Nietzsche we know so well), but if free will is indeed an illusion, then the appropriate response is to move *beyond* accusation altogether. What Nietzsche aimed at was:

> "...no longer resisting anyone or anything, neither the evil nor the evil-doer - love as the sole, as the last possibility of life. [...] "not to defend oneself, not to grow angry, not to make responsible."[xxi]

> "I do not want to accuse; I do not even want to accuse those who accuse. Looking away shall be my only negation."[xxii]

Hard determinists are not in the business of voicing accusations. They should look away. Or, actually, they shouldn't do anything. It is only that *if* they do engage in normative accusations, *we* are entitled to look away. Those 'shoulds' or 'should nots' that they profess do not follow from their position (but perhaps from a different source like consequentialism).

6 CONCLUSION

At root, the problem is that hard determinists cannot help themselves to a normative position without contradiction, not even to a distinct moral position of their own. As soon as they say that a person ought to do or believe something, or that a person ought not to do or not to believe something, they cannot maintain hard determinism. Why not? Because if they say that a person *P* ought to do X and then stick to the conviction that '*P* ought to do X implies that *P* can do X', they should conclude per *Modus Ponens* that *P* is free to do X.

But if determinists next observe that *P* does not do X, as they say *P* ought to do, they will conclude, first, that *P* cannot do X, and next, per *Modus Tollens* that it is not the case that *P* ought to do X. And this makes it once again clear why it is wrong for hard determinists to attribute innocence to criminals. How can one be innocent if there is nothing one ought to do, or refrain from. Of *what* would one be innocent (or guilty)? Of crimes? Again, if we define crimes as actions that ought not to be done (and how else should we?), then according to hard determinism criminals are either free not to commit their crimes or there are no such actions as crimes.

Hard determinism, we conclude, has no moral force. And if it has no moral force, its truth or falsehood has no relevance for moral argument.

[i] In a recent volume of Ratio Iuris, Vincent Geeraerts calls the retributivist perspective the dominant view regarding the justification of punishment. Vincent Geeraerts, "The Enduring Permanence of the Basic Principle of Retribution." Ratio Juris (2022): 1-22. https://doi.org/10.1111/raju.12330

[ii] Galen Strawson, "The impossibility of Moral Responsibility", Philosophical Studies 75 (1994): pp. 5-24.

[iii] Pereboom calls himself a 'hard incompatibilist', thereby implying that the truth of indeterminism (the falsehood of determinism) would also exclude moral responsibility (as desert). See: Derk Pereboom, Free Will, Agency, and Meaning in Life (New York: Oxford University Press, 2014), and Derk Pereboom, "What makes the Free Will Debate Substantive?" The Journal of Ethics 23(3) (2019): pp.257-264

[iv] Smilansky has directly defended his position against Pereboom, see Saul Smilansky, "Pereboom on Punishment: Funishment, Innocence, Motivation and Other Difficulties", Criminal Law and Philosophy 11 (2017): pp. 591-603

[v] Psychological determinism, thus conceived, opens up the possibility of reductionist forms of determinism. If, apparently, our first mental state was caused by something that was not a mental state then why not assume that all our subsequent mental states, including the present one, are caused not by earlies ones but also by something else than a mental state.

[vi] Strawson, "The Impossibility of Moral Responsibility", p. 5.

[vii] Strawson, "The Impossibility of Moral Responsibility", p. 16.

[viii] Smilansky, "Hard Determinism and Punishment", p. 359.

[ix] Smilansky, "Hard Determinism and Punishment", p. 360.

[x] It might be a 'good thing' that virtuous doctors get praise, but not because such doctors deserve it. We will get to the consequentialist way out for hard determinists in the next paragraph.

[xi] It has been argued that such a practice of randomly punishing the innocent (that is, those who did not commit an offense) would not be beneficial. See: Charles Goodman, "Why Would Two-Level Consequentialists Punish Only the Guilty?" Criminal Justice Ethics, 36(2) (2017)" pp. 185-6. This, however, is not what we maintain here. It would indeed be a bad idea to loosen or even sever the link between offense and punishment (how else to prevent crime?). The point, however, is that the justification of punishment is entirely forward-looking (so you can brutally 'set an example': "If you do this, then this will happen!").

[xii] Smilansky, "Hard Determinism and Punishment", p. 362-3.

[xiii] If it is the distinct moral position of hard determinism, as Smilansky is saying here, that we mat not punish criminals to further social interests per se, then it is no longer clear what argumentative work is done by the impending collapse of society upon introducing funishment. Accordingly we may also not fine innocent drunken drivers who violate speed limits, in order to keep the death toll of traffic under control (an estimated 1.3 million fatalities each year worldwide). From funishing traffic offenders, not writing them tickets or compensating them if we do (what would that amount to?), no collapse of society is to be expected, just many thousands of additional innocent victims. Some may consider this result as a redcutio ad adsurdum in itself, but it would not be the 'practical' reductio that Smilansky has in mind. Correspondingly we mat ask: if hard determinism morally requires funishment, and if funishment will lead to the collapse of society, so what? If not Smilansky's practical reductio already a reference to an undesirable consequence of hard determinism: the collapse of society by funishment?

[xiv] Perhaps consequentialism is not the only morality that can do without the concepts of responsibility and desert. Perhaps determinists who still find consequentialism unpalatable, might want to resort to *Animal Ethics* for criminals, and for humanity in general. After all, determinists must regard humans as highly, though not morally, developed animals – as determined as all animals are. And though it is true, as is argued by Peter Singer, Animal Liberation (New York: Harper Collins, 2001) and Jeremy Bentham, "An Introduction to the Principles of Morals and Legislation", Utilitarianism and On Liberty, edited by Mary Warnock (Malden MA: Blackwell, 2003), that consequentialists, at least utilitarians, should be concerned about animal welfare, it is not immediately clear that the reverse is also true: that anyone concerned about animal welfare should be utilitarian, that is: a *maximizer* of aggregate welfare. But what will still essentially drop out of *any* moral principle that might be

derived from Animal Ethics, for *Homo Sapiens* is the reference to innocence and desert. It may be concluded that criminals, as highly developed animals, should be treater in certain ways and not in other ways (for instance not in cruel or monstrous ways), but if the treatment is indeed justified by the distinct moral position of determinism, criminals can have no claim to compensation for their treatment, or a claim to funishment. And so, if determinists decide that Higher Animal Ethics rather than consequentialism suits them as their distinctive moral position, society need not collapse under the weight of funishment.

[xv] This is not to say, of course, that consequentialists might not have a problem with justifying punishment. See, for example David Wood, "Punishment: Consequentialism", *Philosophy Compass* 5/6 (2010): pp. 455-469.

[xvi] Here is a most recent statement by Pereboom: "As a hard incompatibilist I am not agnostic about whether attributions of basic desert blameworthiness are trye of us; I maintain that they are not true of us. But I have maintained, and continue to maintain, that we are not in an epistemic position to rule out that our actual blaming practices are justified on consequentialist or contractualist grounds." Pereboom, "What Makes the Free Will Debate Substantive?", p. 258

[xvii] It is only in responding to a certain critical question "Can we then no longer punish offenders?" and in thereby resisting a position that is clearly compatible with their view ("Que sera, sera." Or, "anything goes.") that they help themselves to a normative position that can provide a different ground for punishment (namely non-retributive). But they forget that any position can be taken, literally 'without offence', from the point of view where they are standing.

[xviii] Peter S. Strawson, "Freedom and Resentment", In: Freedom and Resentment and Other Essays (London: Routledge, 2008), p. 23.

[xix] Jan Verplaetse, Zonder vrije wil: Een filosofisch essay over verantwoordelijkheid (Amsterdam: Uitgeverij Nieuwezijds, 2010).

[xx] Nonetheless there may be a quantum of solace in his response. For this, see Susan Wolf who argues that being *determined* by the True and the Good (and we would add: the Syllogistic) is what it *means* to be free. Susan Wolf, "Asymmetrical Freedom", The Journal of Philosophy 7 (1980), 00. 161-166.

[xxi] Friedrich Nietzsche, "The Anti-Christ", in R. Hollingdale (transl.), Twilight of the Idols/The Anti-Christ (London, Penguin, 1990), p. 30.

[xxii] Friedrich Nietzsche, The Gay Science, translated by W. Kaufman (New York: Vintage, 1974), p. 276.

Free will: a minority report

Arjun Devanesan

In this paper I argue that not only is free will compatible with determinism, but that determinism is in fact, irrelevant to the debate about free will. I start by examining Dennett's compatibilism and his dispute with Gregg Caruso (Dennett and Caruso 2018) over whether people can be held morally responsible for their actions. I argue that the debate actually turns on a failure to disambiguate a concept of control with a concept of causation. Once you accept that causation and control are distinct, there is no link between determinism and free will.

To do this I first examine what I take to be the most prominent account of causation, and the one that prima facie takes causation and control to be related: Woodward's (2003) interventionist account of causation. According to this account (roughly), A causes B iff manipulation (or control) of A leads to a corresponding change in B. Some might be tempted to interpret this account as proposing a necessary connection between causation and control. However, if this is the case, then interventionist accounts of causation are circular - causation appears on both sides of the biconditional. Alternative approaches might take control to simply be a counterfactual condition (Lewis 1979) in which case causation and control are logically distinct. Or at least, control is not reducible to causation.

Next, I examine another reason to distinguish causation and control. Even if we accept that control is an inherently causal concept, we may still argue that it is not reducible to one. It is not uncommon in the sciences to examine the cause of certain phenomena and not have to start at the Big Bang because we think that there is a principled distinction between proximal and ultimate causes. We might even go so far as to think that only rational agents are capable of control without generating much controversy. If so, one might reasonably assert that control is a proximal cause of action that only occurs in rational agents.

Caruso was agnostic about determinism "since indeterminate events are no more within our control than determined ones" (Dennett and Caruso 2018) but his free-will skepticism stems from his belief that our actions are beyond our control. This, I argue stems from a mistaken belief that there is nothing to control beyond causation. But, if one distinguishes control and causation, it is easy to see that the locus of control and the ultimate causal locus for any particular action can be distinguished with the former being personal even if the latter is external. It remains an open question as to what conditions are required for free will, but my suggestion is that it is not determinism or the nature of freedom that is the issue but the nature of control.

References

Daniel C. Dennett & Gregg D. Caruso (2018). Just Deserts: Can we be held morally responsible for our actions? Yes, says Daniel Dennett. No, says Gregg Caruso. *Aeon* 1 (Oct. 4):1-20.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow, Noûs, 13: 455–76.

Woodward, James (2003). Making things happen: a theory of causal explanation. New York: Oxford University Press.

Free Will, Reasons-Responsiveness and Artificial Intelligence

Ville V. Kokko[1]

Abstract. This paper presents a refinement of Christian List's "compatibilist libertarian" view of free will and applies it to the question of AI alignment: In what ways do different artificial intelligences match this definition, and what can this tell us about how to make sure they do not turn against us in small or large ways?

I start off by examining List's view of free will and decision-making and analysing a problem in them. The higher-level indeterminism he proposes runs into the usual problems with indeterminism in free will, and it does not really match the kind of theories of decision-making List evokes to justify it. I continue by suggesting an amendment to the theory where the higher-level indeterminism is replaced by a combination of a deterministic choice function and a non-deterministic set of options. As this is not yet quite enough, I make some further observations about how we need to add a requirement of a certain kind of responsiveness to reasons to describe a free agent.

Having done all this, I present how this theory can be applied to examine the functioning and development of different machine learning and artificial intelligence systems. Even if one does not accept the model as one that appropriately describes human free will, it is useful for evaluating such systems as partial agents and their goal-directed behaviour, and it gives new perspectives on the alignment problem. I look at how some current problems in training artificial intelligence systems are related to how they lack some properties of free agents, at how projected future fears of misaligned AI are related to properties of free agency, and how all this might help us in thinking forward towards solving these problems.

1 INTRODUCTION

In this paper, I present a further elaboration of Christian List's theory of compatibilist libertarianism and then show how the view of free will and agency implied by this view can be applied to understanding what goes wrong when artificial intelligences do undesirable things. My background is in the philosophy of free will, and my exploration of AI here is just one of the first steps I am taking in that direction. Nevertheless, I can already say several things about the alignment problem in the view of this certain concept of freedom and agency on an abstract level.

In the second section, I introduce List's idea of indeterminism on a higher level as a solution to the question of how free will can meet both requirements seeming to require determinism (control) and indeterminism (open alternatives). The basic idea is based on how determinism and indeterminism can appear on different levels of the system. I show how this leaves List with the usual problems of indeterministic free will: control is still compromised in his model. However, I also present a further elaboration of how the model can preserve both a form of open alternatives and the right kind of control by dividing the higher level into a deterministic and indeterministic part.

Next, I argue that something more is needed to complement this model, because it is not clear which part of the person's motivational system should be counted as part of their free decision making and which should be counted as limitations of it. I argue that this can be resolved by looking at the system of the person's interests as a whole, such that those motivations or influences that endanger the harmony of this system by overriding concerns that are more important in the bigger picture are seen as restricting freedom. I also argue that the requirements of alternative choices and control can both be seen as forms of reasons-responsiveness under this model, albeit not matching any of the more specific forms of reasons-responsiveness introduced by John martin Fischer and Mark Ravizza.

In the next section, I turn to the alignment problem and discuss how different existing and hypothetical examples of AIs doing something undesirable can all be explained in terms of the AIs failing to be free agents according to the model presented here, though often, they do not resemble unfree humans so much as not agents at all. Arguably, a real agent must be able to weigh multiple different goals or values at the same time. Thus, it may be desirable to give AIs the ability to do this in order to make them both more competent and more aligned with human goals. Conversely, AIs that are competent actors may automatically have some of these capacities. This may also lead to it being possible to hold them responsible for their actions, or at least usefully treating them as if they are. However, these considerations are very abstract, and more work needs to be done to connect them with the reality of AIs.

2 LIST'S HIGHER-LEVEL INDETERMINISM AND ITS PROBLEM

Christian List's "compatibilist libertarianism" is based on compatibilism with determinism on the lowest level combined with the indeterminism required for libertarianism on the higher level. I will explain the idea here briefly.

List provides a good characterisation of what is meant by levels himself, so I will quote a part of it:

> Facts about the world—in general, not just in relation to human agency and free will—are stratified into levels. In the sciences,

[1] Dept. of Philosophy, University of Turku, Finland. Email: vvkokk@utu.fi.

> "levels" correspond to different ways we describe the world. To make sense of the basic laws of nature, for example, we employ fundamental physical descriptions, drawing on our best theories of physics. We use concepts such as particles, fields, and forces, and specify a variety of equations that describe how physical systems evolve over time. By contrast, to make sense of chemical or biological phenomena, we need to go beyond fundamental physics. Molecules, cells, and organisms all display patterns and regularities that can only be captured at another level of description, using a conceptual repertoire distinct from that of fundamental physics. As philosophers of science have pointed out, even a property as simple as acidity cannot be satisfactorily described in fundamental physical terms alone. In the case of acidity, there is no easily describable configuration of fundamental physical properties that exactly matches this chemical property. In particular, there is no "translation scheme" that fully translates talk of acidity into talk of particles, fields, and forces alone. We need the concepts and categories of chemistry to talk about acidity. And we are here still dealing with a fairly basic example, compared to other, more complex chemical or biological phenomena. [1]

The levels can be seen as ontological or only pertaining to our descriptions of the world, but we do not need to make that distinction here. The important thing is to know that the world can be described partly truly in terms of different such levels. In the case of List's idea of determinism and indeterminism in the case of free will, what matters is that the higher psychological or intentional level on which we describe human actions, the choices may be indeterministic even if the lower physical level on which everything is built is deterministic. To put it briefly, this is because information is lost coming from the lower level to the higher level. Higher-level states can be multiply realised on the lower level, which means that even if you know the higher-level state exactly in the higher-level terms, you do not know the exact lower-level state that realises it. For example, you could know someone's mental state exactly as a mental state but not know the exact molecular state in their brain that corresponds to it. Determinism on the lower level implies that given the exact state of the world on the lower level, only one future is possible, but since the higher-level state does not tell us the exact state on the lower level, then even if the world is determined on the lower level, there may be several possible futures given only the higher-level state.

Levels of description and how determinism and indeterminism can appear on different levels are an interesting and important topic in their own right (one good source on this is [2]; for more on List's characterisation of them, see [3]), but here, it will suffice to know that the same world can be (or at least appear) to be deterministic on some levels and indeterministic on others. This leaves us with List's suggestion that higher-level indeterminism can be enough for what free will requires. Obviously, this does not satisfy all libertarian requirements, nor is it even at all libertarianism as it is normally understood. Instead of trying to answer that challenge, I will here highlight another problem and propose a solution that answers to that. [2]

There is an important problem that faces indeterministic accounts of free will. I will not here try to show its generality, even though that is something I think has not been taken seriously enough on the whole,[3] but I will present how it is a problem in List's account in particular. If human choices are indeterministic on the higher psychological or intentional level, this means that they are undetermined by reasons. List's view holds that every choice that is free must be indeterministic. Thus, if a choice is free, then even if the agent realises that they have very good reasons to choose one option rather than another, this realisation cannot guarantee that the better option gets chosen. Thus, you could realise that one option is better, but freedom implies that even given that realisation, you might make the worse choice. This is practically problematic, nor does it seem intuitively correct.

3 A SOLUTION: MAKING THE CHOICE FUNCTION DETERMINISTIC

List also expresses his view of the required indeterminism on the higher level by saying that different alternatives must be open to the agent, and he asserts that social sciences and theories of decision-making assume indeterminism in this sense, because they hold options to be open and address the question of how an agent chooses between them. [1], [6] To address the problem that control seems to be lost, he and Wlodek Rabinowicz propose that in addition to the set of open options, an agent would have an *endorsement function* that determines which options the agent can do *with endorsement* based on the agent's own reasons. [7] Thus, the set of options that are agentially possible is different from the set of options that the agent can take with endorsement. Only those choices that can be made with endorsement can be made freely. List and Rabinowicz assert that this answers two different intuitions about free will, those being the openness of alternative possibilities and endorsement. Even if we do not think so much of intuitions, these requirements make practical sense for us to be able to make choices in a reasonable sense.

Something is still missing from this solution for it to work. The endorsement function would presumably be deterministic (this is not made clear), and if so, we could avoid the possibility that an agent would freely choose an option that they have reason to reject. This is because if the agent did choose an option they did not endorse, they would definitionally not be choosing freely. However, the endorsement function is not making the choice itself deterministic, since free choices need to be both endorsed and undetermined. Thus, we are left with an odd situation: if you can only endorse some options, then a free choice is one in which you *happened* to pick one of the options that you can endorse, but you also *might not have*, since the choice had to be indeterministic.

This hardly seems right either. It is a strange requirement for a free choice that it have a property that in itself means it might have become unfree. Furthermore, it is always going to be practically problematic and counter-intuitive that we might choose the worse option for no reason even if we know the reasons not to choose it.

My solution at this point is to take the idea of the endorsement function one step further. We can call this a *choice function*, and though we need not here know how exactly it works, we can say

[2] I look at the problem and solution in more detail in a paper that I have so far published as a preprint. [4] Sections 2, 3 and 4 in this paper all contain elements explained in more detail there.

[3] I have examined this in detail in [5], which also explains the sense of *determinism* and *indeterminism* that I assume here. This article is in Finnish, but I am working on an English translation that I will post on my Academia.edu page when it is finished.

it is a function that outputs a choice selecting one out of the set of agentially open options as the choice that is actually made. To avoid the problem of it always being possible to choose the worse option, the choice function can be deterministic. This means that given the options and the reasons for choosing from among them, then if there is a single best choice, that choice will be made. Thus, while the set of options is itself open and does not constrain the choice, and on a level of description only describing the choice function indeterminism is true, we also do not have to worry about possibilities being *so* open that we may choose the worse one for no reason.

This is also what the purported indeterminism in sciences of human behaviour that List describes (and as seen in the kind of models he and colleagues discuss in [7] and [8]) really means. Options are open for agents to choose between, yes, but we do not stop at that and assume they choose randomly. We try to formulate a model that explains which thing they choose or why they choose one thing rather than another.

I am sure it is correct that, for us to choose freely, there must be a set of options that are open in some sense. Examples where options are closed include those where they are unattainable for external reasons, but also ones in which they are that due to psychological reasons. If a person is psychologically unable to do something, then it makes no sense to hold them responsible for it, and freedom and responsibility are usually seen as connected. At the same time, it is necessary that we be able to have control in the sense that we are able to act on those reasons we see that we have. This model with the deterministic choice function and indeterministic set of options combines determinism and indeterminism on the higher level to achieve a version of both of these. This is also a reason for calling it an explanation of free will. However, even if one does not accept that this describes "real" free will, these observations about how agency works can be usefully applied to explain the capabilities and deficiencies of AIs. Before getting to that, I still need to do some more groundwork about how to understand agency in the next section.

4 REASONS-RESPONSIVENESS AND CHOOSING THE RIGHT REASONS

When we are speaking of the choice function and the open set of options, what should be noticed is that it is at least not completely obvious what factors should determine what goes into each of these models. We can certainly look at some examples that are arguably obvious. For example, if addiction makes a person unable to choose to stop indulging in their addiction in a given situation, we have reason not to include the option of stopping in the set of possible options – and thus, we can say that the person has no free choice about whether to stop.

If we do this, we could rely on the general assumption that some things are external to our wills, and some are internal to them. Thus, a strong desire to act on one's moral principles might be considered as internal to the will, and thus acceptable to place in the choice function. This is what List and Rabinowicz do with their endorsement function when they take the example of Martin Luther, who seemingly said he could not do otherwise when he chose to stand up for his principles. [7] (The problem in their solution based on endorsement is shown by their explanation that Luther effectively meant that he could do other thing with endorsement. If his act was indeterministic on the higher level, as per List's compatibilist libertarianism, this would mean that Luther might have done otherwise in spite of his strong conviction, though he would then have felt he could not endorse what he was doing. If we want to say it was not a possibility all things considered that Luther, against his strong convictions, might have done otherwise, we should use the deterministic choice function I am presenting here.) However, the theory at hand can be developed further to explain *why* some internal factors are considered as part of the agent's motives and some other internal factors can be considered as limiting it. The answer does not rely on finding fundamentally different metaphysical categories but in finding distinctions that make a difference based on the nature of agency.

To get started with this, we can note that both the determinism of the choice function and the openness (indeterminism) of the set of options can be explained as serving the same end of *reasons-responsiveness*. The set of options needs to be open for the choice function to choose from so that, whichever of the options there turns out to be a reason to choose, it may be chosen. At the same time, the choice function needs to be deterministic so that if there is a reason to choose one option rather than another, then the reasoned option will really get chosen.

Reasons-responsiveness has notably been used by John Martin Fischer and Mark Ravizza as a condition for responsibility, if not so much freedom. [9] There is no space to present their theory here, so I will only say that I am using the concept somewhat differently. I am not using any of their analyses of weak, strong or moderate reasons-responsiveness, which would not suit the current purposes (due to their all-or-nothing definitions among other reasons). Instead, you can think of reasons-responsiveness in this context as the quality of being responsive to the right reasons, which can be more or less present. The upshot is that the more an agent reacts to the rights reasons appropriately, the more reasons-responsive that agent is. One example of such an idea from earlier in history comes from Thomas Reid, who wrote that God as the ideal agent will always choose based on the best reasons – which does not compromise God's freedom – and that human agents also approach this ideal by being the better determined by the right reasons the more competent they are as agents. [10]

This still leaves with the original question of what the difference is between someone who is reliably self-determined by the right reasons and someone who is un-freely determined by a factor such as addiction or compulsion. However, the roots of my answer to this question are also included in this idea of reasons-responsiveness. We do not need to posit some reasons to be more valuable, right, or part of the self from the start to come to the conclusion that some motives are wrong in the sense of bad for the agent if they override other motives.

The reason for saying this has overlap with Mary Midgley's reasoning for why morality is (as it were) structurally necessary for human beings, which she attributes originally to Charles Darwin. [11] The basic idea is that humans are beings with many conflicting interests, importantly including brief but strong momentary impulses and milder but long-lasting motives. For such a being, it would be self-destructive to follow whatever impulse is strongest at the moment; consider, for example, the impulse to punch someone in anger to all the reasons why you might regret that later. For this reason, we need to make judgements about what is overall the best thing to do considering all our interests, and these judgements need to have an overriding quality compared to our desires at the moment. Even if we did not

have momentary strong impulses, the same basic reasoning would hold with respect to how we can balance all our different interests so as not to undermine others while pursuing one. This is a sort of internalised version of the rationale for morality that we all need to agree to some rules in a society where different individuals have different interests; the different interests within the person play a role similar to the different individuals in society.

Given that something like this is true about the nature of agency, it also gives a criterion for saying which motives can be considered the right ones. One does not even need to accept that this is a correct description of what morality is about, only that it describes something necessary for coherent agency. If we say that to be reasons-responsive is just to be responsive to all your reasons at once, that starting point is enough to lead us to reason that what is required is to consider what takes all our different interests taken together best into account, and to that end, we need to make judgements something like how Midgley describes moral ones.

Thus, consider an overwhelming addictive or compulsive desire. It is a factor that overrides your responsiveness to reasons (at least relatively speaking): you might have quite strong enough reasons all things considered to resist the temptation, but the compulsive nature of it makes it impossible or too difficult to act based on these reasons. Compared to this, consider a strong resolution based on your own convictions that you must make a given choice, even if it goes against some other interests of yours. This plausibly represents a judgement, moral or otherwise, that this thing is what is best all things considered, even if some interests or inclinations go against it. Thus, it is reasons-responsive.

Though this is quite a different perspective than the perspective that includes the choice function and the set of options, it can give answers to what should be included in the choice function – which corresponds to the agent's autonomous choice – and which things should be included as potential limitations on options instead of part of the agent's choice. Basically, some things are overriding *despite* the agent's reasons or interests as a whole, while others are overriding *because* of their place in the agent's set of interests.

There are still a few more observations that need to be made about the nature of reasons-responsive agency in this context, but at this point, we can already move on to talking about AIs and use them as examples about those points.

5 WAYS ARTIFICIAL STUPIDITY IS UN-FREE, AND HOW TRUE INTELLIGENCE MIGHT AVOID THEM

The *alignment problem* is the problem of getting AIs to (aim to) do what we want and what is according to our interests instead of something else. (See e.g. [12].) We could separate the problem of getting AIs to do what their designers want them to do from the problem of how to get AIs to behave ethically or bringing out good results to all of us (see e.g. [13]), but here, I will broadly speak of all of this as part of the alignment problem.

There are many ways in which AIs have failed to be aligned in reality and several more in which they have been imagined to possibly do so in the future. These range from current AIs being creatively stupid by finding ways to do what they are rewarded for instead of doing what the reward was supposed to lead them to do; to the possibility of future superintelligent AIs cleverly destroying humanity either because that is what they genuinely want to do, or because they are still being stupid and were accidentally programmed to do that. An example from that simpler end might be an AI playing a computer game in a way that maximizes its score by moving repetitively in place instead of trying to reach the goal.

When an AI starts mindlessly maximising its reward function, we are faced with a partial agent that is incapable of weighing different interests and goals: instead, it has only one. From its own point of view, if it had one, this can make it free, as there is nothing but the one goal, which means no other goal contradicting it either. However, what we were trying to create was a different kind of partial agent: one that had one overriding goal, sure, for example to play a game well, but that would flexibly realise our open-ended idea of what that would mean. This is an inherent problem with proxy measures, as it ends up being the case that the system that tries to be reasons-responsive to reaching the goal we are aiming for is not the AI but the combination of us and the AI we are training. For an AI to conceive of we want in a way that we can agree with, it would need to have this flexible reasons-responsiveness built within itself.

We humans ourselves have difficulties specifying the rules that even we ourselves follow. Giving a an explanation of how we use common words can be difficult even for philosophers (just consider "knowledge"), different theories of morality that seem to be based on very intuitive principles all face strange counterexamples where they seem to endorse things we would not want to endorse, and if we were supposed to decide which rules of morality to give to artificial agents, we would have plenty of different cultural and other systems of norms that we would have to choose from. (On the last point, see e.g. [13], [14].)

Likely, at least one major reason why it is so hard to specify even one agent's moral rules or even exact goals is that real human agents have different potentially contradicting interests, and the more the agent is competent at reaching them, the more it is the case that they are flexibly balancing their different goals in an *ad hoc* fashion. Even trying to reach a single goal needs to be pursued in a flexible manner. Consider an example where an agent's sole goal in the scenario is to reach a flag in the middle of a field. If there are no obstructions, the most effective way to achieve this goal is to move directly towards the flag. If there is a wall directly in the path between the agent and flag, but it does not extend indefinitely to both sides, then moving directly towards the objective will not work, but moving around the wall first and then towards the flag will. If there is a locked door in the way with no way around, then actions other than moving along the surface will be required. What this requires is teleological action involving the possibility of imagining what you need to do to reach the goal, considering all factors that are relevant. You cannot specify a general rule for what you need to do other than a teleological one like "Do whatever is necessary to reach the goal." Meanwhile, all this was while still assuming the agent has only one goal. What if the agent is a real human, and there is someone lying badly injured on the field and in need of help? In almost all cases, this would be a more important thing than reaching the flag, invalidating even the rule to always do whatever it takes to reach the flag.

Thus, while it is hard to give artificial agents instructions, values or training parameters that make them do what we would want them to do in all cases, maybe it is also the case that we cannot make them competent agents that do not trip over their own feet that way anyway. Maybe we cannot create truly intelligent agents without also making them teleological and all-things-

considered reasons-responsive. This might mean that we need a different approach to how we train AIs, or it might mean that if our current approaches turn out to be enough to produce truer intelligence, then those approaches will as a side effect produce intelligences that can understand what we want of them better. [4]

Approaching morality as a continuous open question like this could help with the goal of alignment as Iason Gabriel articulates it: "For the task in front of us is not, as we might first think, to identify the true or correct moral theory and then implement it in machines. Rather, it is to find a way of selecting appropriate principles that is compatible with the fact that we live in a diverse world, where people hold a variety of reasonable and contrasting beliefs about value." [15] I will not say the approach of reasoning flexibly in different situations can simply replace the need for deciding which values AI needs to be programmed to respect (though I do not see that option as completely impossible), but such flexible intelligence might well help in navigating this problem, perhaps even in cooperation with the AI itself. [5]

There is a general line of thought that some AI disaster scenarios are nonsensical because they assume an AI that is both so smart as to be incredibly dangerous and stupid enough to simply do the wrong thing. The line of thought about the nature of agency and intelligent action gives this simple objection some possible more precise content: an agent capable of intelligently achieving its goals in the world would necessarily already have the structural capability to weigh different goals and means flexibly in different situations. Even if it was trained from a starting point of maximising some values, to achieve the *general* intelligence to pursue any goal in different circumstances, it would need to have transcended that.

None of this removes the possibility of an AI *intentionally* acting against humans. It *might* be that what I said above about the nature of morality would imply that AIs capable of functioning as actors would be more likely to have some kind of morality than we might otherwise assume, but the points I made are so general that we cannot yet conclude anything like this. In fact, much of human morality is arguably pretty weird considering this explanation of morality, and at any rate, there is a great variety of different ideas of what is moral or immoral among humans. Still, in grappling with questions of both AI alignment and that of improving AI capabilities, these points are good to keep in mind.

As a final observation, the question of AI becoming reasons-responsive can also have bearing on whether AIs can be treated as morally responsible. Once again, the question as I put it here is not whether they would be genuinely responsible for their actions, but whether it would make practical sense to treat them as such. A necessary requirement for it to make sense to hold an agent responsible and hope to motivate the agent to behave better is for the agent to be able to take the knowledge of being held responsible into account as a reason, whether as a question of expecting to face consequences or as a moral question important in its own right, or both. Thus, such an agent needs to be reasons-responsive in the right way. Making artificial agents free in the sense described in this paper might thus enable us to treat them as responsible, or at least usefully treat them as if they were responsible to motivate them to do the right thing.

8 CONCLUSION & FUTURE WORK

In this paper, I have briefly sketched out a number of observations about explaining free, autonomous choice using a combination of a modified version of Christian List's compatibilist libertarianism, a general concept of reasons-responsiveness, and an idea of harmonising one's different interests in a way that is related to Mary Midgley's view of why morality is necessary for beings like us. These are all topics that could be explored at much more length, as I have done elsewhere [4], [16] and also will in my doctoral dissertation. Then, I have made a start at applying these ideas to look at what is going on with artificial intelligences failing to perform intelligently as well as to the worries we have about future artificial intelligences. This suggests that the reasons why AIs might do things we do not want them to do overlap with reasons why they are not fully functional agents, which might point to some ways forward in improving and aligning AIs. The observations made here are very preliminary, however, and I hope to continue from them in the future using more detailed ideas about how AIs operate.

REFERENCES

[1] C. List, *Why Free Will Is Real*. Cambridge, Massachusetts: Harvard University Press, 2019.

[2] L. Floridi, *The Philosophy of Information*. Oxford: Oxford University Press, 2011.

[3] C. List, 'Levels: Descriptive, Explanatory, and Ontological', *Noûs*, no. 53:4, pp. 852–883, 2019.

[4] V. Kokko, 'How to Amend Christian List's Theory on Free Will to Answer the Challenge from Indeterminism', *Qeios*, 2023, doi: 10.32388/73P1K2.

[5] V. Kokko, 'Satunnaisuusargumentti ei jätä liikkumatilaa inkompatibilismille', *Ajatus*, vol. 81, pp. 29–60, 2024.

[6] C. List, 'Free Will, Determinism, and the Possibility of Doing Otherwise', *Noûs*, no. 48:1, pp. 156–178, 2013.

[7] C. List and W. Rabinowicz, 'Two Intuitions about Free Will: Alternative Possibilities and Intentional Endorsement', *Philosophical Perspectives*, vol. 2014, no. 28, pp. 155–172, 2014.

[8] F. Dietrich and C. List, 'Reason-Based Choice and Context-Dependence: An Explanatory Framework', *Economics and Philosophy*, no. 32, pp. 175–229, 2016.

[9] J. M. Fischer and M. S. J. Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge ; New York ; Melbourne ; Madrid: Cambridge University Press, 2000.

[10] T. Reid, *Essays on the Active Powers of Man*. Edinburgh: Edinburgh University Press, 2010.

[11] M. Midgley, *The Ethical Primate: Humans, Freedom and Morality*. London ; New York: Routledge, 1994.

[12] B. Christian, *The Alignment Problem: Machine Learning and Human Values*. New York: W. W. Norton & Company, 2020.

[13] L. Dung, 'Current cases of AI misalignment and their implications for future risks', *Synthese*, vol. 2023, no. 202:138, pp. 1–23, Oct. 2023.

[14] P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery, 'Human-aligned artificial intelligence is a multiobjective problem', *Ethics Inf Technol*, vol. 2018, no. 20, pp. 27–40, Oct. 2017.

[15] I. Gabriel, 'Artificial Intelligence, Values, and Alignment', *Minds and Machines*, vol. 2020, no. 30, pp. 411–437, Oct. 2020.

[4] Related points to both of these are discussed in [13]: that more advanced AIs may be less likely to be misaligned due to their capacities, and that deep learning methods may be inherently prone to lead to misalignment. Meanwhile, training AIs with multiobjective methods is discussed in [14].

[5] I examine the question of morality as flexible and unfolding in a little more detail in [16] – another paper in Finnish that I plan to translate into English to upload that version on my Academia.edu page.

[16] V. V. Kokko, ‘Mitä “hyvä” tarkoittaa? Arvottavien väitteiden mielekkyydestä’, *Paatos*, Jul. 2018, [Online]. Available: http://www.paatos.fi/2018/07/16/mita-hyva-tarkoittaa-arvottavien-vaitteiden-mielekkyydesta/

Free will as an argument against physicalism

Yaren Duvarci

A robust conception of free will often requires that our thoughts, feelings, and actions originate from something beyond the purely physical processes of the body and brain. This creates a fundamental tension between free will and physicalism. However, free will is rarely used as an argument against physicalism, likely because the exact nature of their conflict has not been clearly explained. This essay addresses this gap by showing the real reasons for this incompatibility. Unlike accounts that focus on determinism at the level of action and alternative possibilities, I argue that the real problem is the disappearing agent in physicalist theories. I maintain that true control hinges on the agent's capacity to be the source of their actions—a capability that physicalist theories fail to account for due to the disappearing agent problem. This analysis, I contend, reveals a fundamental incompatibility between free will and physicalism, suggesting that the existence of free will offers a robust argument against it.

New Approaches, Same Old Problems: Science, Free Will and the *Catch-22 of Libertarianism*

Samuel Michaelides[1]

Abstract. Despite arguably having an intuitive appeal, the libertarian approach to freedom and responsibility is often considered to be incoherent due to its necessary focus on undetermined acts and its seemingly contradictory requirement for a sufficiently determining agent-control. In the attempt to meet both of these conditions, it is argued by critics that libertarian theorists are led inevitably into what I have called the *Catch-22 of Libertarianism*; with the meeting of one, precluding the meeting of the other. The position continues to have its supporters in philosophy, however, and more recently a number of scientists have also joined the debate on the side of libertarianism. In this paper, I consider two models developed by scientists – Peter Ulric Tse's Criterial Causation theory and Roger Penrose and Stuart Hameroff's Orchestrated Objective Reduction theory – and asses each against my *Criteria for Coherence* to see if they can be considered intelligible. Finding that both fail to meet my criteria, I move on to consider why they have failed and what is it about the libertarian approach that makes developing a coherent model so difficult. I then discuss the potential irrelevance of both indeterminism and determinism when it comes to the free will debate, before finishing with some closing remarks about what adopting an "irrelevantist" stance might mean for other, non-human, agents including a suitably advanced AI.

1. Introduction

Libertarian theories of freedom and responsibility argue that a "true" freedom of the will – one which allows an agent to take advantage of genuine alternative possibilities and yet retain ultimate control over her decisions and actions – *does exist* but is incompatible with the deterministic picture that the future is causally necessitated by the past. Such theories have always been subject to varying levels of criticism, mistrust and even ridicule, and the attacks have not just come from philosophers, but also from many in the scientific community who often favour a compatibilist approach (if they subscribe to free will at all). However, some researchers working in fields such as neuroscience, physics, and mathematics *have* taken up the challenge to produce a coherent libertarian theory, likely spurred on by seemingly indeterministic developments in fundamental physics, famous experiments in neuroscience and psychology (such as those by Benjamin Libet and his followers), a growing understanding of the structure and function of the brain, and also advancements in fields such as artificial intelligence. The question is whether their contributions fare any better than those of their philosophical counterparts or whether the oft-cited problems – such as incoherence, inescapable arbitrariness, and plain mysteriousness – are just fundamental to any indeterminist model.

In what follows, I attempt to answer this question and ultimately argue that the libertarian models presented, though certainly inventive, *cannot* be considered fully coherent as they fail to meet the stated goals of the libertarian project in a consistent and non-contradictory fashion. I go on to discuss the reasons for this failure and the implications of it for both the particularly reductive, scientific approaches considered, as well as for libertarian theories in general. I argue that it is in fact the explicit incorporation of indeterminism itself that causes all the problems and that its inclusion in this fashion is not *actually necessary* in order to meet the libertarians' aims and objectives. I finish by questioning whether notions such as determinism and indeterminism should really be seen as entirely irrelevant to the issue of whether we have free will or not and consider what this might mean for other non-human and even non-biological agents, such as an advanced artificial intelligence.

1.1 The Libertarian Challenge

The goal for the libertarian theorist is to reconcile their understanding of free will with indeterminism: meaning they need to provide a positive account for how an agent can act freely and with sufficient control in an indeterministic universe. There are, of course, numerous compatibilist arguments which attempt to show that freedom can be reconciled with *determinism* but, for libertarians, the version of freedom which the compatibilists are arguing for is normally not considered to be the "Gold Standard" they themselves are seeking and believe exists. The perceived lack in a deterministic universe of what has been called "deep openness" by Alfred Mele (Mele 2014, p.2) – the ability to act differently given the exact same past conditions – leads to accusations that compatibilist agents are limited only to exercising "freedom of action" and thus do not possess a true "freedom of the *will*." The former (somewhat broadly defined) I take to be the ability of the agent to act freely (in the manner of their choosing) without suffering any proximal impediments or constraints, such as being physically coerced, under threat of violence or under the influence of drugs and so forth. The latter (again, broadly considered) I take to add the further condition that the agent not be (wholly) subject to any form of necessitating causes over which they can have no control, whether that be the motions of atoms or the mechanical output of their own cognitive processing. Thus, to attain the Gold Standard freedom of the will, an agent requires the ability to carry out (at least some) actions or make (at least some) decisions that are not wholly determined by prior states of the universe, but which are nonetheless sufficiently within their control.

Many retort that this libertarian notion of "free will" is simply nonsensical and argue that any attempt to formulate a theory based on such principles can only end in failure for reasons which generally include the following:

1. Decisions or actions which are *not* wholly determined by past conditions in conjunction with natural laws can inevitably only be the result of *random* processes and thus can never be truly under the control of the agent.

[1] Recent PhD graduate, University of Hertfordshire: sammichic@icloud.com

2. Any attempts by libertarians to ascribe a sufficient form of control to the agent by firmly rooting their decisions and actions within a prevailing character (or *will*) presumably must involve some level of determination by this character, but this then raises questions over the formation of that character and whether the agent themselves can actually be responsible for forming it.
3. Attempts to address points 1 and 2 (as well as counter other criticisms and objections) usually leave libertarians trying to account for free will by invoking obscure or mysterious forms of agency and/or causation which, as well as often being difficult to comprehend, can bring with them ontological commitments that are difficult to accept.

Point 1 is a particularly common attack levelled at libertarian theories and leads to the argument that the required indeterminism – far from providing the agent with a necessary, unrestricted universal framework within which she can act freely, responsibly and with control – is actually destructive and serves only to *diminish* her control because choices are ultimately either determined or they are not, and undetermined choices, by definition, cannot truly be under the control of anyone or anything.

In regards to point 2, it would indeed seem that if we are to make any sense of the notion of a decision or action being fully within an agent's control – of a choice being ascribed to them in a manner in which we would consider them to be responsible for it – such decisions must somehow be firmly rooted within that agent's character or, to be more specific, must be the product of reasoning processes derived from their accumulated mental content (the sum of their nature, experiences, learning, desires, beliefs, proclivities and so forth), which I have termed their *Current Established Psychology* (CEP). However, introducing an essentially determining psychology (even if considered as only a "local" incidence of determination which is in no way indicative of any wider deterministic framework) in order to provide a sufficient form of agential control inevitably raises questions concerning where the responsibility lies for the formation of such psychology; setting off a seemingly infinite regress of a CEP formed by decisions made by a previous version, which was in turn formed by decisions made by a previous version, and so on back to a point before any rational decisions could possibly be made.

So, (according to the critics of the libertarian position) the positing of decisions and actions undetermined by past conditions leaves such decisions and actions overly subject to random factors which cannot be under the control of the agent, thus negating genuine agential responsibility. Any additional positing of some form of determining CEP in order to stave off any accusations of randomness likely leads to a problematic regress concerning the formation of the character from which an agent's decisions and actions would flow, which *also* seems to negate any true agential responsibility. Any further attempts to halt such a regress by specifically utilising the indeterministic foundations to posit causal breaks at specific points in the various chains of reasonings, decisions and actions just reintroduces the random factors, *once again* negating any true agential responsibility.

This dilemma speaks to the heart of the coherence question and is what I have referred to as the ***Catch-22 of Libertarianism***.

1.2 Two Key Objections

Two specific objections have been formulated out of the more general criticisms above, designed to refute all libertarian theories – the Luck Objection and the Basic Argument. Both objections are key to the question of coherence, and both will therefore need to be countered if we are to consider the theories and models put forward by the scientists to be successful.

The Luck Objection

The Luck Objection is born out of one of the key tenets of libertarian theories – often labelled the "Alternative Possibilities" (AP) condition – which refers to the requirement that, at any point in time, an agent must be free to *act* or *act otherwise* whatever the past conditions and natural laws. Or, to put it another way, an agent must have been able to have acted otherwise at a particular point in time, *all past circumstances being exactly the same*. This principle is seen by libertarians as essential to prevent an agent's decisions and actions from being wholly determined by antecedent events; such determination, they believe, being incompatible with their Gold Standard freedom of the will (and this being their predominant reason for rejecting compatibilism). The problems arise when we consider the potential implications of this AP condition: that if indeterminism means agents can genuinely have different alternatives open to them that all still remain consistent with their past and the laws of nature, it could be argued that we cannot conclusively account for why one alternative is chosen over another. This leads some to the conclusion that the final choice can therefore only be a matter of luck.

The Basic Argument

The second key objection is born out of one of the other main tenets of libertarian theories of free will – the requirement for "Ultimate Responsibility" (UR). In short, the UR condition states that if an agent is to be ultimately responsible for any act, she must also be responsible for any preceding actions, decisions or events which can be said to be causally responsible for that act. Or, to put it another way: if an agent is to be considered fully responsible for acts dictated by her *will* (by her character or CEP), she must be fully responsible for the *formation* of the will from which such acts flow. The problems which arise in consideration of this condition have resulted in many believing free will (in the libertarian sense) to be simply impossible *whatever the truth of the fundamental nature of the universe, be it deterministic or indeterministic*. Such a view is encapsulated in the Basic Argument, an old argument recently formulated in the most developed way by Galen Strawson,[2] which focusses on the (proposedly incoherent) notion that a human being could be *causa sui*: the cause of oneself. The overriding claim of this argument is that we cannot ultimately choose what we do because we cannot ultimately choose *who we are*. We may act freely – be able to carry out "free actions" – in the sense (as discussed above) that we are not constrained by any obvious external forces, such as

[2] See Strawson, G. 1986, 1994 and 2002.

mental health conditions, addictions, coercive captors and so forth, but if we are not ultimately responsible for the formation of our CEP from which our actions arise then we still would not have the Gold Standard freedom of the *will* that libertarians are after.

Utilising my own terminology, the argument proceeds as follows. We act as we do, in the present, due to our current established psychology (CEP) – the way we are now, mentally speaking – as developed by our heredity, choices and experiences. If we are to be genuinely responsible for how we act in the present, we must be ultimately responsible for the *formation* of our CEP. But, in order to be ultimately responsible for the formation of our CEP, we would have to have carried out some actions in the past for which we can be held responsible which contribute to the formation of our CEP. But, in order to be responsible for *those* actions we must have been responsible for the formation of our established psychology (EP) that was in operation at that time in the past. But, in order to be responsible for the formation of our EP that was in operation at that time in the past, we would need to have been acting responsibly at an earlier time still, which would contribute to the formation of this past EP. And thus begins a regress. So free will is impossible because, ultimately, we cannot be the cause of ourselves.

1.3 The *Criteria for Coherence*

The above discussion highlights the arguably intractable problems for libertarian theories of freedom and responsibility. On the one hand, the Luck Objection appears to show that any posited breaks in the causal chains to prevent our decisions and actions being sufficiently caused by antecedent conditions leads to there being an unacceptable amount of arbitrariness present in the process. On the other hand, the Basic Argument claims to show that the positing of any kind of determining control would seemingly entail a determining psychology (CEP) that we could not possibly be responsible for creating, resulting in us not being ultimately responsible for those present decisions. Hence the *Catch-22 of Libertarianism*.

The key question is, then: exactly what *would* it take for a libertarian theory to be considered fully coherent? By which I mean, how could a theory successfully meet *all* the stated goals of the libertarian project in a consistent and non-contradictory fashion. In order to try and answer this question, I have elsewhere developed my *Criteria for Coherence*[3] which includes six principles that identify the goals libertarians wish to achieve and which helps to crystallise the issues that seem to be at the root of any apparent incoherence.[4]

It proceeds as follows:

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.

P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

It is this *Criteria* that I will now use to assess the coherence of two libertarian models produced by scientists.

2. Peter Tse's Criterial Causation

Neuroscientist Peter Ulric Tse claims to have identified a particular brain mechanism which he believes means free actions *can* navigate the seemingly elusive "middle path" between the unfree extremes of deterministic inevitability and indeterministic randomness. If proven correct, it seems he is potentially offering a way to achieve the libertarian Gold Standard freedom of the will as well as a way *out* of the *Catch-22 of Libertarianism*. According to Tse, the reason many researchers (in both science and philosophy) have been driven towards viewing indeterminist free will as an illusion is down to a misapprehension of how neurons in our brains encode and transmit information. The traditional picture of neuronal activity as solely a succession of action potentials (also known as *spikes*) cascading through neural circuits – with one spike able only to trigger another in a cause-and-effect, feedforward, linear fashion – gives the impression of our brains as deterministic biophysical machines. Tse argues, however, that this image does not represent the whole story and such spikes can actually alter how other neurons in the circuit will respond to future inputs without necessarily triggering them immediately – like "changing the combination on a padlock without opening it." (Tse 2013a, p.28)

On Tse's model, our brains respond to situations requiring decisions or actions by setting up relevant criteria in neurons and neural circuits, which will then need to be met by subsequent inputs in order for the neurons to fire. If we then combine the brain's use of this mechanism to set up criteria for future firing (which Tse has termed "Criterial Causation") with neuronal inputs arriving at the synapses that are fundamentally indeterministic in nature due to the amplification of events at the quantum level – and which will either cause the neurons to fire or not (perhaps leading to future firing and/or further re-setting of criteria) – *and* have it all playing out within the arena of an active and

[3] In the interests of clarity, the full appellation *Criteria for Coherence* is intended to be read in the singular, denoting the full set of principles as a whole, as is the single word *Criteria* when used as shorthand for the full appellation. At such times, italics will be used to avoid any confusion. For more extensive arguments concerning the development of the *Criteria* and to witness its deployment against libertarian theories developed by various philosophers, see Michaelides 2024.

[4] I make no claim to huge originality with the individual principles proposed here, only that their drafting together to form a list of conditions against which libertarian theories can be formally assessed is a useful step.

coordinating consciousness, we have a model for human decision-making and action which realises a particular form of top-down mental causation that is not completely deterministic nor completely random: thus meeting both the Alternative Possibilities (AP) condition and the Ultimate Responsibility (UR) condition and delivering the Gold Standard in a coherent form.

> [C]riterial causation offers a path toward a strong free will that passes between the unfree "Scylla" of determinism and the equally unfree "Charybdis" of randomness. (Tse 2013 p.22)

Tse wishes to debunk the notion that – due it being apparently underpinned by entirely deterministic and mechanical neural processes – there can be no genuine mental causation (and thus no causally efficacious *will*) and to propose a model that *does* offer an agent sufficient control over her actions. He outlines his own conditions which must be met in order for an agent to realise what he terms a "strong free will":

> We must have (a) multiple courses of physical or mental behavior open to us; (b) we must really be able to choose among them; (c) we must be or must have been able to have chosen otherwise once we have chosen a course of behavior; and (d) the choice must not be dictated by randomness alone, but by us. (Tse 2013, p.133)

These conditions, it will be noted, align closely with my own *Criteria for Coherence* and thus Tse's intention is to hold his libertarian theory up to a similar standard.

2.1 Criterial Causation, Indeterminism and Consciousness

There is essentially three parts to Tse's theory: the brains ability to carry out neural recoding via criterial causation; microscopic indeterminacies in the synapse which lead to unpredictable inputs; and internally manipulable conscious thoughts. I shall briefly consider each part in turn.

Criterial Causation

As just discussed, criterial causation relies on a particular conception of our neuronal biology. A neuron up close would appear a rather blunt instrument; an organic switch operating mindlessly in connection with linear sequences of electrical and chemical signals. The activity of each neuron at first glance appears wholly (and dumbly) reliant on the inputs from those that precede it, creating a mechanical and seemingly deterministic picture of human action reaching causally back into the past. Tse argues that, on this picture, the notion that higher-order mental events associated with what we would likely label as conscious reasoning could alter the very physical events on which they are based would be unacceptably circular. Indeed, he argues, if we accept this traditional view of neuronal workings, we must accept that the driving force behind all decisions and subsequent behaviour is wholly the mechanistic action of the stimulus-response, feedforward, all-or-nothing neurons and any kind of higher-order mental causation of the sort Tse believes would be required for free will is simply ruled out.

It would seem there are no immediately obvious causal gaps which a libertarian theorist of free will could perhaps exploit, however Tse believes suitable gaps *do* in fact arise, generated by the way in which synaptic connections can be rapidly altered in a coordinated fashion by neurons themselves. This is because, according to him, neurons, rather than solely responding crudely to stimuli from other neurons, can actually dynamically alter their synaptic properties and those of the neurons around them to pre-set criteria that will determine whether or not an action potential is generated when an input next arrives at the synapse. Tse is attempting to change the perception of neurons as solely feed-forward linear classifiers – largely passive receptors – whose firing is decided exclusively by whether or not the summation of the electrical potential of the incoming stimuli passes a fixed threshold at the axon hillock; presenting them more as active players in cognitive processing.

The scientific explanation for exactly how our neurons might achieve this (biologically speaking) is expectedly complicated, (see Tse 2013 for the detail), but for our purposes here it is enough to recognise that there is good evidence to suggest that it is at least a possibility that changeable properties of synapses could be altered in this way to recode the inputs that will make a neuron fire in the future, and these changes can be made without triggering an action potential in the affected neurons and can happen, if necessary, extremely rapidly. Tse's criterial causation is therefore presented as a biological mechanism that he believes can be employed to rapidly reshape the circuitry of the brain (presumably at the behest of some form of central processing/decision making module – CEP) and thus be used to implement a volitional will. As such, it is a key component of his overall model of mental causation and his theory of free will.

The Role of Indeterminism

Tse argues extensively for indeterminism as the ontological reality of our universe[5] because, although he believes the criterial causation part of his theory would, in isolation, still hold up in a deterministic universe, he does not believe genuine free will is compatible with determinism due to the libertarian view that this would result in a failure to meet the AP condition. On Tse's account, it is the Indeterminism at the microscopic level of neuronal inputs and the inherent variability this creates, coupled with an agent's ability to set her own neural criteria, that underpins free will as it allows her to retain volitional control over her actions whilst still being able to "do otherwise." Criterial causation pre-sets criteria in neural circuits to guide/restrict output, but it does not *determine* the output and Tse is clear that were we to run the sequence of events over again we would not necessarily get precisely the same outcome.

Tse gives the following example as an illustration of his view:

> If I ask you to think of a politician, your brain sets the appropriate criterion in neurons involved in retrieval of information held in your memory. Perhaps Margaret Thatcher comes to mind. If it were possible to rewind the universe, you might think of Barack Obama this time, because he also meets the criterion. This process is not utterly random, because the answer had to be a politician.

[5] See Tse 2013, Appendix 1, p.241.

> However, it is also not deterministic, because it could have turned out otherwise. (Tse 2013a, p.28)

Tse is therefore required to argue for a biological foundation which would allow indeterminism at the microscopic level to have an effect at the macroscopic level, which he does in great detail in his book. Once again, it is not necessary to go into the complex neuroscience here but is enough to acknowledge that there is evidence to suggest quantum level indeterminacies could be playing a role in the transfer and binding of neurotransmitter molecules and thus affecting the outcomes of our neural mechanics.

The Role of Consciousness

Tse's reasons for arguing that consciousness must play a key role in mental causation, and therefore in free will, is because *without* a conscious agent with intentional states (rooted in a CEP) that could put her criterial causation into action volitionally to attain goals, he believes human beings could be considered mindless zombies.

> The brain of a zombie who lacked consciousness could use this mechanism too, but we would not say it had free will. To have free will requires that our self – that which we feel directs our attention around our conscious experience – has some say in the matter of what we do or think. If consciousness plays no part in the synaptic reweighting process, there is hardly a free will worth having. (Tse 2013a, p.28)

As to the precise nature of the role consciousness plays, Tse argues it is most likely to be in providing an essential common format for relevant executive operations in the brain, such as endogenous attention, which assess potential behaviours and/or thoughts involved in fulfilling desires. In fact, endogenous attention, according to Tse, is *required* for internal deliberative processes and consciousness is *required* for endogenous attention; thus, internal deliberation can only happen at all where consciousness is in play.

Tse is proceeding from an assumption that we have experiential states and without them we would undoubtedly act differently. Experience and consciousness, therefore, must play some key role in the causation of our actions (at both the mental and neuronal level) and this role – according to Tse – is providing a format for the assessment of various possible courses of action and allowing the internal attentional manipulation of the associated thoughts. This then leads to the neuronal recoding required for his criterial causation. Exactly why Tse believes that such operations can *only* take place consciously is not completely clear. He appears to argue somewhat by stipulation; stating there simply must be a level where competing courses of action can be assessed against the overarching goals of the agent. He thus maintains that there is an important link between endogenous attention, working memory and experience.

So, although an important part of the process, consciousness alone is not sufficiently causal. Rather, consciousness allows for the endogenous attentional operations that reach the decisions and it is these processes that then go on to impose the relevant criteria at the neuronal level. Consciousness, then (it would seem), is a form of *necessary arena* in which deliberation can occur and decisions can be made. Once these decisions are made, the framework for realising them is rapidly encoded onto neural circuits. The ultimate outcome is then generated in a partially random fashion by the arrival of variable inputs.

Tying it all together

If we bring together all three parts of Tse's theory we can attempt to build it into a "real-world" example. A young athlete, Jessica, is driving to an important qualifying race when she witnesses an elderly gentleman collapse in the road ahead and has to decide whether to stop and help him and risk missing her race or just make her turn and drive on. We can imagine her paused in the street, with her hand on the car door handle, internally debating whether to jump out and help or leave him to his fate. On Tse's model, we might imagine that upon seeing the elderly man fall, the relevant neurons from her sensory organs convey information about the situation to the cortical areas responsible for assessment, evaluation, and planning operations. Such operations likely involve endogenous attention – which necessarily requires the framework of consciousness – in order to engage in a deliberative reasoning process. Let us say that, at the culmination of such processes, a decision is made to stop and help. Signals are then sent out from the initial planning regions to other brain areas involved in the next processing stage which encode into the relevant neural circuits (via synaptic re-weighting) "stop and help."[6] Once this criterial pre-setting is achieved, indeterministically variable inputs arrive at the circuit and either meet the criteria or not. If it is met, the circuit is activated in such a way as to produce an appropriate plan of action and further signals are sent out (now in a presumably determined fashion) via the motor neurons to put the chosen plan into action.

There seem to be a number of problems and confusions with this picture. A key issue is that arguably the most important part of the process – the reaching of the decision to either stop and help or drive away – is left largely unexamined and unexplained and appears to occur *prior* to the enactment of the very criterial causation model Tse is proposing. A secondary problem is the level of randomness present in the theory in regard to the indeterministically variable inputs which may or may not meet the specified criteria. As we shall see in the next section, these issues, rather than defeating them, arguably do more to *feed* both the Luck Objection and the Basic Argument.

2.2 The Luck Objection and the Basic Argument

The Luck Objection

[6] Perhaps, more specifically, Tse would wish to argue that what is actually encoded is something like "decide how to stop and help." The impetus to help is there but the variability is in how best to achieve the goal. This might fit better with his example of encoding a thought about a politician but then allowing variability for who is chosen. But if this is correct then in situations such as Jessica's the key decision to help is made deterministically and the only variability is in *how* to help, which would not seem to advance the libertarian case. Part of the issue with Tse's account is he is not clear on points of detail such as this.

Tse argues that, despite his emphasis on the importance of indeterminism, luck or randomness is eliminated from the important parts of his model by the pre-setting of the criteria which essentially dictate what neural outputs are possible. In combining criterial causation with indeterministically variable inputs, we now have a biological mechanism for the setting up of a framework within which a desired outcome could be achieved (or at least some version of it), but which will not determine one outcome over another and which could thus have led to a different outcome. But if no particular outcome is apparently being *directly selected by the agent*, can the agent be said to actually be in control of her actions? If the active control Tse is arguing for reduces to merely the encoding of criteria in order to achieve a desired outcome – the actual fulfilment of which is then reliant on indeterministically variable inputs which may not realise such an outcome – can we truly argue such a model would meet the Gold Standard? If the variable inputs are the ultimate arbiters of our actions, does this not result in complete randomness?

In truth, if we were to accept Tse's model, we would seem to be left with a rather limited sort of freedom at best – and seemingly *not* the Gold Standard we are trying to achieve. Applied to Jessica's situation, it would seem that what would be encoded into her neural circuits if she decided to assist the fallen man would be "stop and help" and then variable inputs would arrive which would dictate the nature of precisely *how* the help is given. But if variable inputs could lead to actions as different as, say, getting out and performing CPR or simply stopping the car and calling an ambulance, it seems valid to consider whether they could equally lead to her driving away altogether; apparently reinstating the problem of luck. Even if this is not a real possibility and once the instruction "stop and help" is firmly encoded into her neural circuitry the only variation is the precise nature of how that help is delivered, the key decision to help is made *before* any process of criterial causation-plus-indeterminism takes place. We appear to have an agent causally implementing a desire but no explanation for how this desire itself could be formulated in a way that would not be considered either the product of deterministic neural processes or a completely random arising. As Tse is clear in his wish to avoid randomness in key areas of his model, the desire must presumably arise deterministically from Jessica's CEP, but then the precise way her desire is to be implemented is apparently decided by the random process involving variable inputs. This would, therefore, not appear to be the "middle way" Tse is claiming to have found. In reality, what we have is an initial deterministic process which is then followed immediately by an indeterministic one, and adding two things together is not the same thing as finding a middle way between them.

The Basic Argument

Tse specifically targets Galen Strawson's Basic Argument in his work because he believes that his theory, with its focus on changing the criteria for *future* firing rather than attempting to show how mental events could alter their own physical basis in the present, is particularly well positioned to avoid the related traps of self-causation, circularity, and infinite regress (and other problems associated with the ultimate responsibility (UR) condition) that the argument highlights.

Tse's strategy is to claim that Strawson's argument simply does not apply to his theory because criterial causation does not involve the apparently contradictory process of self-causation at all. Mental events, Tse argues, supervene on their physical substrates and thus for a mental event to alter the very physical substrate on which it supervenes would be tantamount to a form of self-causation and would be logically impossible. However, in his theory, mental events do not need to alter their own physical basis in order to be sufficiently efficacious. What they do is set up criteria for *future* physical activity.

Whilst Tse's approach could be considered quite original, it is difficult to see how his line of reasoning does any damage to Strawson's argument. All one has to do in order to reassert the consequences of the Basic Argument is to once again point to the omission from Tse's model of any examination of the process by which our brains "come to the decision" in the first place, before the criterial encoding takes place. In order for an agent to reach a decision based on sensory stimuli at T1, and thus set the appropriate criteria at T2, she must do so from within the parameters of her established psychology *as it is at T1* for it to be considered as *her* decision. And this, according to the Basic Argument, can only be a consequence of previous decisions or actions occurring at T0, T-1, T-2 and so on. All the mental work that the Basic Argument is focused on, therefore, occurs *before* Tse's criterial causation comes into play.

On Tse's model, either the main decision is made deterministically from within this standpoint (from the agent's CEP) prior to the encoding of criteria such as "stop and help," in which case the decision to intervene or not is effectively determined by circumstances beyond the agent's control, or the main decision is not made prior to the criterial encoding and what is encoded is as generic as "decide whether to intervene or not," in which case the outcome is entirely random and dependent on the indeterministically variable inputs, and again beyond the agent's control. Placing great emphasis (as Tse does) on the fact that in his model criteria are being set up in the *present* to influence *future* firing also does not really gain him anything. Positing a supervening (yet entirely physical) mental event capable of influencing future physical/mental/informational events and claiming this solves the problems the Basic Argument raises just does not follow. The actual decisions for what such mental events wish to make happen in the future remain subject to the ever-problematic *Catch-22 of Libertarianism.*

2.3 The *Criteria*

In summary, Tse has argued that it is possible to have a libertarian free will that allows for undetermined but self-selected actions. Neurons are not simply blunt stimulus-response devices, solely at the mercy of their presynaptic inputs, but have the ability to recode their own response to stimuli (and that of other neurons around them) altering the informational criteria for when they will or will not fire. Patterns of neuronal activity harnessing this ability can therefore be causal in a way over and above the physical basis in which such patterns are ultimately realised and it is in this way that an agent can exert a form of top-down mental causation. Information, albeit realised in physical activity, can be causal of future information that will be realised in future physical activity. When we add indeterminism into the picture so that the inputs

received are inherently variable and unpredictable, and include the ability to consciously manipulate the contents of our endogenous attentional operations in deciding what criteria to encode in the first place, we have, according to Tse, all the ingredients we need for a strong (Gold Standard) free will.

I will now assess Tse's theory against my *Criteria for Coherence*. The first two principles can be taken together.

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.

Tse's theory is indeed (overall) an incompatibilist theory. Though he states the criterial causation part of his model would add biological weight to any compatibilist theory, he believes the addition of indeterminism is required to allow for a strong free will and thus argues extensively for indeterminism as the likely state of the universe, and incorporates it prominently into his theory. Thus, I would argue the first two principles have been met.

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

Initially, it might seem clear that Tse's theory also conforms to the AP condition as it states that, although criteria are preset in advance of the arrival of the inputs, the inputs themselves are unavoidably variable due to indeterminacies being influential at the synaptic level and thus an agent could choose a particular course of action but then, were the universe to be wound back, could chose differently. Tse's example of thinking of a politician would seem to conform to this as a thought of Margaret Thatcher could just as easily be a though about Barack Obama. However, in other instances it seems more specificity is being encoded by the criteria, which puts complete conformation with this principle in doubt. As we discussed when applying Tse's model to our Jessica example, if the only undetermined variation available to her is *how* to intervene, not *whether* to intervene, it could be argued this is only a very limited form of available alternatives and not in keeping with the spirit of the AP condition. Thus, it is not clear that this principle has been met.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

Tse would certainly argue that this criterion has been met, but it is not clear this is actually the case. Tse's model potentially sheds light on how an agent can exert some form of top-down control over their neural architecture in order to realise desires and attain goals through criterial encoding, but such encoding (as argued above) is presumably enacted *post-decision*. The decision first needs to be made and needs to be based in the agent's CEP in order for it not to be considered solely arbitrary. There is no discussion concerning exactly how the decision/desire is arrived at in the first place, leaving us again with such a state being either deterministically generated or arising randomly. Plus, the inclusion of indeterminism by way of random neuronal inputs arriving at preconfigured synapses brings into question precisely *how much* control Tse's model is providing the agent in any case. Ultimately, with Tse's model apparently featuring deterministic structuring followed by an undetermined arbitrating input, it could be argued that his model for decision-making is actually *part* determined and *part* arbitrary. Thus, I cannot say this principle has been met.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

Tse's general answer to the problem of self-causation and regress is to rely on the notion of his criteria affecting *future* as opposed to present neuronal firing, in conjunction with the (at least partial) causal break posited by the prominent role of undetermined inputs. This, however, does not remove the question of the seemingly inevitable regress created in the actual decision-making or desire formation process. As we have seen in our discussion of Tse's model, the problem just shifts to the relevant processes occurring *before* criterial causation takes place. As such, principle five has not been met.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

Though in many ways rigorously scientific, there are a number of assumptions made and concepts included which would likely be considered contentious by some and Tse's apparent reliance on emergent phenomena, causally efficacious higher-level patterns of activity, and a supervening active consciousness could be considered unnecessary invocations of mystery. In addition, Tse's arguably ambiguous usage of terms such as "informational" and "mental" lend further obscurity to his model, resulting in his meeting of this principle being cast into doubt.

In conclusion, as judged against the *Criteria*, Tse's model for the Gold Standard libertarian free will cannot be considered coherent. Whilst an interesting neurological model for how our brains might operate, Tse's combination of criterial causation and indeterministic inputs produces a process of mental causation which, instead of being *neither* determined nor arbitrary, manages in reality to be a combination of *both*. In addition, despite claiming he is presenting a theory that provides for a "strong free will" Tse's theory really only presents a mechanistic procedure for the *implementation* of an agent's desires (more, free *action* than free *will*), with little or no discussion concerning key parts of her decision-making process and how such a process could avoid the *Catch-22 of Libertarianism*. Lastly, Tse's equivocations concerning the mental and physical aspects of human ontology and lack of clarity over key terms such as "information" or "informational" when discussing neural states adds an unnecessary level of obscurity to his model. Ultimately, Tse sets himself the task of meeting almost the same criteria by which I have judged his approach and has failed to meet both sets.

3. Roger Penrose and Stuart Hameroff's Orchestrated Objective Reduction

Anaesthesiologist Stuart Hameroff's paper "How quantum brain biology can rescue conscious free will" (2012) is based on his Orchestrated Objective Reduction (Orch OR) theory of consciousness, which he has developed with the mathematician and physicist Sir Roger Penrose. It is Hameroff's contention that, in their attempt to explain the origin and place of consciousness in the universe, he and Penrose have developed a theory that can not only offer a mechanism by which consciousness is generated (and show why it should not be considered epiphenomenal) but which can also account for the autonomous causal agency required for the Gold Standard freedom of the will. Consciousness, they argue, involves non-algorithmic processes which prevent its activity being solely deterministic and thus their model can provide the conscious agent with a mechanism by which they can have a direct controlling role over the biological workings of their own brain. As Hameroff states in the paper's abstract: "…Orch Or can account for real-time conscious causal agency, avoiding the need for consciousness to be seen as epiphenomenal illusion. Orch OR can rescue conscious free will." (Hameroff 2012, p.1)

3.1 OR, Orch OR, Microtubules and Temporal Manipulation

Objective Reduction (OR) is Penrose's own theory for the mechanism by which a quantum state that is evolving unitarily as a superposition of possible states in line with Schrödinger's wave equation, reduces to a single classical reality: often referred to as the "collapse of the wavefunction".[7] OR is in competition with other interpretations such as the Copenhagen Interpretation, Many Worlds theory or Hidden Variables theory and states that reduction occurs once an objective physical threshold of gravitational (space-time) separation has been reached.

Penrose's OR is suitably detailed and complicated, but it contains two elements which are key for our purposes. Firstly, it involves at least some processes which Penrose argues are entirely non-computational or non-algorithmic (the terms computational and algorithmic are used interchangeably, essentially to mean anything that can be simulated on a computer).[8] Secondly, each reduction of the quantum state is accompanied by what is described as a discrete moment of "proto-consciousness" – an element of experience and awareness which, although in its most ubiquitous form is considered primitive and non-cognitive, has the potential to be harnessed within a wider scheme that ultimately results in full human consciousness and causal control of behaviour.

As discussed previously when considering the work of Peter Tse, the prevailing view of how our brains operate has tended to focus on the brute actions of the neurons, which then often leads to accusations that the entire process is solely algorithmic in nature and any emergent consciousness would likely be epiphenomenal and non-efficacious; which then further raises difficult questions concerning free will. Tse's attempt to address this issue was to focus on the capacity of neurons to reorganise themselves in order to implement desired behaviours and not just passively respond to inputs. Penrose and Hameroff, who also view the algorithmic conception of neural processes as a threat to free will, go somewhat deeper into the physical makeup of the neurons in order to look for gaps in the seemingly ballistic processes which could be exploited – though not in a way that would be considered solely random. As above, the suitable gap in current physics that Penrose and Hameroff have identified (and where they believe entirely new physical theories could be developed) is the seemingly discontinuous process of quantum-state reduction, to which they have added an associated and fundamental element of consciousness. What is needed in order to turn such instances of proto-consciousness into a fully functional working model is a mechanism by which the individual instances can be coordinated into any kind of coherent "whole," which could then act as some form of controlling entity. This is where the "orchestrated" part of Orch OR features.

Prior to any reduction, the quantum state evolves unitarily (as per the Schrödinger equation), which involves the relevant quantum computations associated with the various superposed states. This process then terminates by Penrose's OR once the required gravitational threshold is met. It is this process of quantum computation, prior to reduction, that Penrose and Hameroff argue needs to be "orchestrated" in order for any form of agent control to enter the system but, for this to occur in the brain, the relevant superposed state needs to be maintained for long enough without it spontaneously collapsing due to environmental decoherence (standard ORs would collapse randomly and rapidly preventing any real orchestration). As Tse also discusses in his work, the prospect of quantum effects being relevant in such large structures as neurons and synapses is a contentious issue, with many arguing that such warm and wet conditions as found in a human brain would be a very hostile environment for maintaining fragile quantum states. As such, Hameroff's main contribution to the theory has been to identify a biological location which could provide a more friendly environment in which quantum computations could progress successfully. This location is within the microtubules, which form part of the cytoskeleton of neurons and other cells and perform various different functions.

As with the physics, the biology is detailed and complex, but the important point for our purposes is that there would appear to be some evidence that microtubules could provide a suitable biological environment in the brain, capable of both maintaining quantum superposed states for long enough to allow for any required orchestration (influenced by synaptic inputs, memory and other cognitive functions) that brings in non-computable (but also non-random) elements and is also capable of directing/manipulating synaptic activity once the OR threshold has been met and the state collapses. OR without the orchestration (without the *Orch*) would just be at the mercy of its environment and subject to random elements, but when orchestrated

[7] These processes are notoriously difficult to interpret (and even comprehend) but, in very simple terms, the quantum wavefunction is the evolution over time of the different probabilities associated with quantum states. These different states exist in what is known as a quantum superposition of simultaneous possibilities until a measurement is made, at which point the wavefunction "collapses" into a single reality.

[8] This idea has arisen out of Penrose's arguments against the prospect of a strong artificial intelligence and its perceived algorithmic consequences for human conscious reasoning. For more on this see Penrose 1999 and 1994.

("adequately organized, imbued with cognitive information, and capable of integration…" Hameroff and Penrose 2013, p54) and kept isolated from random environmental influences for long enough Orch OR can occur fully, which results in moments of consciousness and allows for agent control.

> *Consciousness results from discrete physical events; such events have always existed in the universe as non-cognitive, proto-conscious events, these acting as part of precise physical laws not yet fully understood*. Biology evolved a mechanism to orchestrate such events and to couple them to neuronal activity, resulting in meaningful, cognitive, conscious moments and thence also to causal control of behavior. These events are proposed specifically to be moments of quantum state reduction (intrinsic quantum "self-measurement"). Such events need not necessarily be taken as part of current theories of the laws of the universe, but should ultimately be scientifically describable…In the Orch OR theory, these conscious events are terminations of quantum computations in brain microtubules reducing by Diósi–Penrose 'objective reduction' ('OR'), and having experiential qualities. In this view consciousness is an intrinsic feature of the action of the universe. (Hameroff and Penrose 2013, p.39)

Hameroff and Penrose specifically discuss the famous experiments carried out by Benjamin Libet and his followers – which many believe are a real stumbling block for free will – in order to demonstrate the potential benefits of the application of quantum theory to human actions. Libet's experiments, which focussed on recording electrical activity in the brain of their subjects whilst also asking them questions concerning their conscious intentions, seem to show that conscious perceptions occur, in fact, only *after* a decision to act has been taken physiologically, ruling out any conscious causal efficacy and casting doubt on the possibility of genuinely willed action (except perhaps the capability to consciously reverse or overrule a non-conscious impulse to act).

Penrose and Hameroff argue contrary to this that quantum theory shows it is in fact possible to refer quantum information *backwards* in classical time (in their case due to the temporal non-locality caused by OR), which could account for Libet's findings and potentially restore an efficacious consciousness. Subjects in the Libet experiments do not falsely remember acting consciously but, rather: "…a quantum explanation with temporal non-locality and backward time referral enables constructive "filling in" from near future brain activity, allowing real time conscious perception" (Hameroff 2012, p.7). It is in this way, according to Hameroff, that our consciousness remains the true arbiter of our willed actions, retaining its capacity to regulate axonal firings and thus coordinate real-time behaviour even if it is in some sense sending quantum information backwards in classical time to do so.[9]

Jessica and her Time Travelling Consciousness

To summarise (and overly simplify) the Penrose-Hameroff Orch OR theory: quantum activity within microtubules (and via gap junctions) creates a vast network of entangled particles in superposed quantum states, which evolve in line with the Schrödinger equation until the threshold for Penrose's (non-computational, gravitational) OR is reached, at which point the wavefunction collapses, a new and unknown process takes place and a moment of "proto-consciousness" occurs. When such instances are suitably *orchestrated*, these combine to produce the sort of subjective consciousness associated with phenomenal experience and intentionality (integrated with memory) which can then influence neuronal firings/inhibitings (potentially) controlling bodily actions. The backward referral in time of quantum information is proposed to prevent consciousness being rendered epiphenomenal and to prevent this being either a completely algorithmic or a completely random process, thus providing the created consciousness with an efficacious rather than solely an observing role and thereby apparently rescuing free will from both algorithmic inevitability and random environmental influences.

Penrose and Hameroff do not seem to have fully extrapolated their model up to the level of visible human action themselves but, were we to apply it to our athlete Jessica's dilemma, it might proceed as follows: upon seeing the elderly man fall to the ground, perceptual information would filter through Jessica's brain, initiating quantum computational activity within neuronal microtubules which would spread out across relevant regions of her brain through entanglement via gap junctions. The quantum state would be maintained within the isolated environment for a sufficient amount of time for a level of orchestration (Orch) to occur; presumably involving her perceptions, memory, and a variety of other cognitive applications generated via her current established psychology (CEP) which can in some way influence the computation process. At some point after the superposition has been maintained long enough for some kind of agential involvement/guidance, a threshold of sufficient gravitational separation between the superposed quantum states is reached and the quantum computations terminate by Penrose's objective reduction (OR) – a new, unknown, non-algorithmic (but ultimately physical), process accompanied by conscious awareness. The OR reduction (in a sense) *chooses* the classical microtubule state Jessica's brain "collapses" into, which becomes the output of the quantum computation and which then triggers axonal firings which ultimately cause Jessica to, say, exit her vehicle and assist.

Some apparent issues are immediately evident. Firstly, it is not completely clear how the "discrete conscious moment" is related to the objective reduction. At times it is stated that the OR is "accompanied" by a conscious moment, at other times the words used are "associated with," and at other times it appears the conscious moment is *identified with* the OR.[10] This makes it difficult to utilise the concept in any kind of "real-world" example of cognition.

Secondly, there is not a great deal of discussion concerning what is actually occurring during the orchestrated quantum computational process. It is not clear how such a process might be

[9] Hameroff in fact states that Libet himself argued for the possibility of subjective information being referred backward in time in order to save an efficacious consciousness from his own experiments. He states that Libet's assertions were "disbelieved and ridiculed…but never refuted." (Hameroff 2012, p.7)

[10] See Hameroff and Penrose 2013, pages 53, 59, and 67 for examples.

influenced in any way by the agent's CEP, in order to allow for sufficient control and meet the ultimate responsibility (UR) condition. Thirdly, at times the orchestration process itself appears to relate to the quantum computations occurring prior to OR, at other times it appears to relate to the requirement to orchestrate the conscious *results* of the OR, and at other times it seems to relate to a combination of both.[11] Given the vast numbers of neurons involved in even the most basic kind of cognition (not to mention the immense complexity of the quantum phenomena involved), I suspect the orchestration concept would indeed have to apply both to the quantum computation phase and the resulting moments of consciousness, but the lack of clarification on this point is one of the reasons it is a difficult theory to apply at any kind of psychological level.

3.2 The Luck Objection and the Basic Argument

The Luck Objection

Penrose and Hameroff argue that the important addition of a layer of "orchestration" to their model (allowed for by the maintenance of the quantum-superposed state for a sufficient length of time) prevents the process from being overwhelmed by random factors, as they state it *would* be were the state to decohere quickly due to entanglement with its environment. Indeed, they state that the vast majority of OR events are *not* orchestrated as part of some coherent structure so would be overwhelmed by random factors and, as such, would have no significant conscious experience associated with them. These would only produce primitive "proto-conscious" events (without information or meaning) which, on their own, only form the isolated ingredients for a fully functional and experiential consciousness. When they *are* orchestrated, however, the randomness disappears. What is not clear, though, is just *how* this is achieved. It is argued that quantum computations progress within the microtubules, after which OR (which likely requires some non-computational new physics to fully explain it) acts to "select" specific states. But what precisely is guiding such a selection, is not discussed in any detail.

A further problem with assessing the Orch OR theory against the Luck Objection is that it is also not clear from the presentation of the theory exactly what parts of the process may or may not be considered to be deterministic or indeterministic. At some points in the literature, it is suggested that the quantum computation process that occurs prior to OR proceeds entirely deterministically in line with the Schrödinger equation. At other points it is suggested that cognitive influences during this computation could add an important "x-factor," suggesting the process may not be entirely determined. If there is to be any indeterministic influence, it would seem that the most likely stage for its involvement would be during the OR process but, if this is the case, then this raises familiar issues concerning how such a process could come under the agent's control. If the CEP's involvement is essentially *determining* then Orch OR does not seem to escape from the algorithmic/deterministic issues its authors have identified as a threat to free will (relating both to the influence of the CEP over the decision-making process and the formation of the CEP itself). However, if the involvement of the CEP is *not* determining then we are left with the question of what then would "tip the balance" in favour of one choice over the other. The inclusion of the concept of backwards time referral, potentially utilised by consciousness to send quantum information back to influence the quantum computations occurring prior to OR, does not appear to solve the problems of agent-control as there would remain a point (or an interval) at which conscious decisions would have to be made prior to the information being referred back and all the issues raised above concerning determinism or randomness could just be applied to that stage in the process.

If the moment of consciousness is *identified with* OR, rather than just accompanying it in some disjointed sense (as the literature at some points seems to suggest),[12] then perhaps it could be said that the (CEP influenced) quantum computation terminates in a conscious moment which acts as the moment of "decision," the OR counterpart of which then selects the appropriate microtubule states which initiate the relevant axonal firings, and so forth. The question then becomes whether the resulting conscious decision is "informed" or "necessitated" by the preceding quantum computation. If it is not necessitated by it, and thus the decision could equally have been different, then the Luck Objection appears to become valid once again.

Rather than answer the Luck Objection, the Orch OR model (at least as extrapolated up to the level of psychology here) seems likely to contain a good deal of random elements and disconnection between key phases of the model and any attempts to remove either leads back to the algorithmic/deterministic consequences Penrose and Hameroff wish to avoid.

The Basic Argument

The question here is whether the Orch OR model can break any impending regress in such a way that the agent could retain a satisfactory control over her actions and thus meet the UR condition. As discussed above, much of the debate rests on exactly where in the quantum/cognitive processes described any causal breaks might come and also at what stage a decision is likely to be made, neither of which are clearly identified.

If the decision is essentially made during the quantum computation process – as orchestrated by the CEP – then the issue of regress arises in connection with both the unitary evolution of the quantum state *and* the formation over time of the CEP. If the decision is made by the conscious moment, then this is either fully informed by the previous computation or finalised at that moment in conjunction with the agent's CEP (the development of which can be traced back through the life-history of the agent); both scenarios again leading to a regress. If the conscious moment makes the decision and refers information backwards in time, the regress remains due to the required involvement from the CEP prior to the information being sent back. Granted the regress issue is slightly confused due to the convoluted timeline, but the principle remains.

[11] See Hameroff and Penrose 2013, pages 59, 71 and 73 for examples.
[12] See Hameroff and Penrose 2013, p.67.

As with the Luck Objection, the Orch OR model does not appear to be able to offer any new avenues of argument against (or response to) the Basic Argument.

3.3 The *Criteria*

I will now review the Orch OR theory against my *Criteria for Coherence*, starting (as previously) with the first two principles taken together.

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.

Sir Roger Penrose himself, it has to be said, at times appears noncommittal about both the existence of free will and whether or not the universe will turn out to be deterministic or indeterministic. Hameroff, though, is much more explicit in this area, using Penrose's arguments against a purely computational consciousness as a basis to dismiss physical determinism and epiphenomenalism (the existence of both – or either – of which he believes would preclude free will) and claiming that the Orch OR theory can account for a conscious free will that is clearly of a libertarian variety. I would argue, therefore that both principle one and principle two have been met.

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

Again, Penrose and Hameroff would likely point to the fact that their model rests on a quantum-mechanical foundation (which inherently contains indeterministic/probabilistic elements) and also point to the non-computational elements of the proposed quantum-mechanical consciousness as evidence of agential freedom from any form of determinism; all of which means the model arguably meets the AP condition. Principle three is therefore also met.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

Though they do not really get into the specifics of decision making or elevate their model to the level of human psychology in any great detail, Hameroff has stated clearly that he believes the Orch OR theory does provide a model for conscious choices which is neither random nor determined and fully under the control of the agent. It is argued that the model provides for an efficacious consciousness which can exert direct physical influence within the brain in a manner which is not solely the result of algorithmic processes and neither subject to entirely random environmental factors. However (as we have seen above), the model seems in fact to contain some processes which appear deterministic and some which appear essentially random. When considered against the Luck Objection the model could not account for how an agent could have full control over their decision to implement a particular course of action while resting on the indeterministic foundations of quantum mechanics. Principle four has therefore not been met.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

As with potential elements of randomness, various potential regresses also appear prevalent in the Orch OR model. Wherever the main locus of agent-causal control is placed – be it during the CEP orchestrated quantum computations or the OR/Consciousness event – if the act is not to be entirely disconnected from its antecedent motivations, some form of problematic regress appears inevitable. The addition of the notion of consciousness referring quantum information backwards in classical time to in some way influence events does not help matters as it just shifts the faculty subject to the regress to a different part of the process. Principle five, therefore, is also not met.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

While Penrose and Hameroff would likely wish to distance themselves from the word "mystery" when presenting their theory, there are certainly (by their own admission) plenty of "unknowns." This is perhaps understandable given that the subject matter focuses on providing a quantum-mechanical/neurological foundation for consciousness (the amalgamation of two subjects notoriously mysterious in many ways) but it could be argued that more work needs to be done in certain areas before the theory can really be put to test, philosophically speaking. The mysteries in their theory, they would say, are not *necessarily* mysterious but rather simply scientific facts waiting to be discovered. That being said, I still do not believe principle six has been met.

In conclusion, the Orch OR theory presents an intriguing model of how quantum mechanical operations in the brain could provide a foundation for cognitive processing and conscious control. At its core is no doubt solid scientific reasoning concerning Penrose's Objective Reduction as a competitor to other interpretations of the measurement problem in quantum theory and concerning Hameroff's belief that microtubules would provide a favourable environment for quantum computation in the brain. The problems come when these central notions are used to construct a framework for not only consciousness but also free action. It is at this stage the model becomes too vague in many of its assertions, and this becomes particularly problematic when attempts are made to fully extrapolate the framework to the psychological level in order for it to be considered as a basis of a libertarian theory of free will.

The theory raises a host of questions as many of its key elements are unclear, including exactly what kind of entity consciousness is in this model, how or when the orchestration process occurs, what kind of psychological elements are likely to be involved, how they would be involved, whether the orchestration process is a deterministic process, whether the OR is a discontinuous event, how the orchestration process influences OR, how or why such an

event results in moments of consciousness, and so forth. But the most crucial thing to note for our purposes is that even if we were to decide on answers for such questions, it seems apparent the theory would suffer from all the same problems that have beset other libertarian theories. In the face of the Luck Objection, the theory cannot account for how either the preceding quantum computations or the OR/Consciousness event (wherever the locus of agent control might lie) could exert a satisfactorily non-determining control over an agent's actions. In the face of the Basic Argument, the model could not account for how an agent could exert a non-random control anywhere within the quantum mechanical framework in such a way as not to lead to an infinite regress. Lastly, the theory failed to adequately meet the *Criteria* and so cannot be considered coherent.

4. Conclusion

The purpose of applying my *Criteria for Coherence* to the scientific theories considered in this paper was to test whether either of them have managed to successfully provide a model for free action which can meet all the true ideals of the libertarian position in a coherent manner – something I have argued elsewhere the philosophers have failed to do.[13] Ultimately, I have concluded that neither manages adequately to meet all six principles and thus, by this measure, cannot be considered coherent. This then leads to the question of exactly *why* they have failed, and what are the implications of this failure.

4.1 The *Catch-22* and the Root of the Problem

When considering what, precisely, is it about certain principles contained within the *Criteria* that makes adhering to them so problematic, it becomes apparent that the root cause of the failure of the considered approaches is, in fact, the very thing that makes them libertarian theories – the sacrosanct requirement to include indeterminism explicitly in their models.

The researchers whose work I have considered here seem quite clearly to assume (as do philosophers supportive of the libertarian position) that the only way principles 3, 4 and 5 can be met is by basing their models within an explicitly indeterminist framework (hence the addition of principles 1 and 2), and this assumption can ultimately be traced to their belief that neither the Alternative Possibility (AP) condition nor the Ultimate Responsibility (UR) condition can be met in a deterministic universe. This then leaves them with the difficult (perhaps impossible) task highlighted by principles 4 and 5 of needing to develop a model for action which manages to conjoin two necessary, but seemingly contradictory faculties: undetermined decision-making and a sufficiently determining agential control. What we have seen, is that neither of models presented manages to achieve this in any coherent fashion. As the *Catch-22 of Libertarianism* highlights, whenever a theorist tries to evade the Luck Objection and introduce some kind of determining control to reduce the level of apparent arbitrariness in their model (which arises due to this necessary adherence to the AP condition and the indeterminism libertarians are convinced such adherence entails) they increase the likelihood of a vicious regress being instigated, and whenever they try to evade the Basic Argument and introduce more causal breakages in order to reduce the danger of a regress (and thus adhere to the UR condition, for which libertarians again believe indeterminism is essential) they inevitably increase the level of arbitrariness. Often, in such attempts to rid their theories of all arbitrariness *and* a problematic regress, they just end up with models in which their agents are overly subject to *both*.

The question we need then ask is "why should this be the case?", and one answer could be that achieving the level of agent-control libertarians themselves claim they require in order to attain their Gold Standard freedom of the will is just *not possible* whilst adhering to the main (indeterminist) tenets of their own libertarian position. There just seems to be no libertarian form of control which can match the arguably more straightforward, antecedently determining control afforded to such alternative (deterministic) philosophical positions as compatibilism. Again, we are left to consider why this is so, and the answer strongly suggesting itself is that the included indeterminism can only ever confuse any picture of agent-control by emphasising seemingly intractable issues concerning the nature of the link between and agent and her actions; a link we intuitively require an intelligible explanation of in order to attribute a level of agential control that could be considered sufficient for any ascription of responsibility.

The *really* important question that comes out of such discussions is: is this absolute focus on indeterminism *required* in order to achieve an acceptable model of free action that meets the stated goals of the libertarian project – and thus offers a sufficient level of freedom – and within which an agent can be said to have full control and be responsible for their deeds? As I have argued elsewhere, it is my contention that it is *not* actually necessary, and both the AP and UR conditions, and all the libertarians' stated goals, can be met regardless of whether the universe should turn out to be indeterministic or not.[14] In trying to use indeterminism as a means to meet both the AP and UR conditions (and thus to meet principles 3, 4 and 5), the scientists whose work I have considered – and whose mechanistic, reductive approach serves to really highlight the problems – have, I believe (like the philosophers before them), fallen into an unresolvable contradictory position, being led down paths that end largely in confusion, incoherence and/or mystery. Instead of being viewed as some essential and active ingredient in libertarian theories, indeterminism should arguably be seen as irrelevant to the accounts of freedom and responsibility such theories produce.

4.2 Irrelevantism, Agency and AI

Given the forgoing discussion, I believe an argument can be made that we can retain all the positive features of the best libertarian theories *whatever the fundamental physical nature of the universe should turn out to be* and can meet all the important goals of the libertarian project in an entirely non-contradictory and coherent fashion by simply removing any explicit reference to indeterminism from the theories. This is something the libertarian

[13] See Michaelides 2024.

[14] See Michaelides 2024, section 4.2. The argument for this is rooted in the notion that the included indeterminism is not actually doing anything conceptually important. It is not, in fact, *doing any positive work at all* and thus all the well-established confusions its inclusion creates could arguably be said to be entirely unnecessary.

models themselves are not able to do specifically *because* their belief that their own criteria *can only* be met within the framework of an indeterministic universe forces them to employ particular strategies to meet them that result in such contradictions. To be clear, arguing that the libertarian belief that indeterminism is categorically *required* to meet their stated goals is mistaken does not mean I am just arguing for compatibilism (at least not in its traditional form). It could perhaps be suggested that – in contrast to the "impossibilists" (such as Galen Strawson) who believe that any genuine freedom of the will must be incompatible with both determinism *and* indeterminism – my position is a "possibilist" approach, which argues that a genuine freedom of the will should be considered to be compatible with *both* potential physical frameworks. But what I am *really* arguing is that whether the universe turns out to be deterministic or indeterministic should, in fact, just be considered as wholly *irrelevant* to the issue of freedom and responsibility – so perhaps we should label me an "irrelevantist."

The point I am ultimately making, is that what the evaluation of these reductive scientific models really shows is that the common practice (in both science and philosophy) of trying to extrapolate upwards from microscopic processes to the level of psychological reasoning simply does not work. Questions concerning fundamental physics, therefore, simply need not be considered as part of the debate on free will.

More generally, I believe, it really is time the debate moved on from this encumbering framework of "Determinism vs Indeterminism" or – more specifically – from the argument over which of the two is *more* compatible with free action and gave serious consideration to this notion that they should both indeed be considered as largely irrelevant to discussions of freedom and responsibility, with neither bearing any relevance to substantive models of agential action and decision-making. The common tendency to view determinism and indeterminism as some kinds of universal forces (often malign) that have the power to *interfere* in the lives of human beings is, I would argue, simply mistaken and it is the historical fetishisation of these scientific notions – which are really nothing more than potential physical frameworks for the motions of particles or waves and can thus be relegated to issues concerning the prediction and calculations within physics – which has arguably hindered progress in the development of a more appropriate conceptual framework which could be applied in order to further the debate on freedom and responsibility; one that focuses more on *agency* as a psychological phenomenon, and one which has its own terms to describe mental events that should not necessarily be seen as automatically, ontologically inferior to those proffered by the physical sciences. Physics, it could well be argued, is simply not in the business of saying anything about agency, which does not figure as a notion in any laws of nature, and, by return, philosophers debating free will have no need to concern themselves with fundamental physics. This does not of course mean that there is no underlying physical universe, connected in some important way to our thoughts and desires, only that such desires cannot be *explained* by invoking fundamental physical processes.

As a final thought, the adoption of an irrelevantist stance – one which argues there is an explanatory separation (if not exactly a physical one) between the biological/mechanical substrate of our neural workings and the level of our psychological reasoning, which renders the microscopic makeup of such a substrate essentially irrelevant to the question of free will and agency – might not only release human-beings from the perennial determinism/indeterminism-free-will-removing-trap, but other life-forms as well, and potentially even a suitably advanced artificial intelligence. With such a gap in play, it seems it could be argued that the precise nature of such a substrate – whether organic or silicone based – itself has no obvious effect on whether an agent would be considered to have free will or not. This then seems to lend additional credence to the notion that an advanced artificial intelligence, based on something like a quantum computer, which is capable of learning, remembering, autonomously generating options and choosing between them would possess basically all the same attributes as a human agent that would likely be considered relevant in a discussion as to whether such an agent has free will. The only differences would perhaps be the notion that the AI was initially externally programmed and also maybe a question over the human brain's capability for plasticity and rapid organic regrowth but, in regard to the former, there is a good argument our genetic inheritance – plus initial conditions – is essentially a form of programming not dissimilar and, in regard to the latter, it does not seem a stretch to imagine the development of some kind of synthetic neuronal-type material which would be capable of similar regrowth and reorganisation.

Whilst I do not have the room to delve deeper into this idea here, one possible takeaway could be that I suspect many of those who support the prospect of an AI being considered to have free will would assume this could only work on a strictly deterministic/compatibilist model. Perhaps what I have discussed here shows this is not necessarily the case, and the real key to understanding agency – whether rooted in an organic life-from or not – is to put the physics to one side and contrate on what being a free agent *really* amounts to.

References

[1] Grush, Rick & Churchland, Patricia Smith. (1995). "Gaps in Penrose's toilings." *Journal of Consciousness Studies* 2 (1):10-29.

[2] Hameroff, Stuart. (2012). "How quantum brain biology can rescue conscious free will". *Frontiers in Integrative Neuroscience*. 6: 93.

[3] Hameroff, Stuart & Penrose, Roger. (1995). What 'gaps'? Reply to Grush and Churchland. *Journal of Consciousness Studies* 2 (2):98-111.

[4] Hameroff, Stuart & Penrose, Roger. (1996). Orchestrated objective reduction of quantum coherence in brain microtubules: The "Orch OR" model for consciousness. *Mathematics and Computers in Simulation* 40:453-480.

[5] Hameroff, Stuart & Penrose, Roger. (2013). "Consciousness in the universe. A review of the 'Orch OR' theory." *Physics of Life Reviews* 11: 39-78.

[6] Kane, Robert, (ed.) (2011). *The Oxford Handbook of Free Will*. New York: Oxford University Press.

[7] Mele, Alfred R. (2014). "Kane, Luck, and Control: Trying to Get by without Too Much Effort." In *Libertarian Free Will: Contemporary Debates*, ed. David Palmer, 37-51. New York: Oxford University Press.

[8] Michaelides, Samuel, (2024). "Is Libertarian Free Will an Inescapably Incoherent Concept?" *PhD Thesis, University Of Hertfordshire archive:* https://uhra.herts.ac.uk/handle/2299/28181

[9] Penrose, Roger. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.
[10] Penrose, Roger. (1999). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.
[11] Strawson, Galen. (1986). *Freedom and Belief*. Oxford, GB: Oxford University Press.
[12] Strawson, Galen. (1994). "The Impossibility of Moral Responsibility." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 75: 5-24.
[13] Strawson, Galen. (2002). "The Bounds of Freedom." In *The Oxford Handbook of Free Will*, ed. Robert Kane, 441-460. New York: Oxford University Press.
[14] Tse, Peter Ulric (2013). *The Neural Basis of Free Will: Criterial Causation*. MIT Press.
[15] Tse, Peter Ulric. (2013a). "Free will unleashed." *New Scientist* Issue 2920.

Nietzsche on free will

Yifan Li

Keywords: Nietzsche, Free Will, Autonomy, Stoicism, Amor Fati

Abstract

Free will has traditionally been conceived as a kind of power to control one's action, either through the ability to have acted otherwise or by being the ultimate source of one's actions. The tension between this conception and determinism has sparked long-standing debates, which P.F. Strawson critiques as overly intellectualised, often losing sight of the practical context. This paper argues that Nietzsche's practical philosophy offers an alternative perspective on free will, one that may shed light on contemporary debates. While most scholars explore Nietzsche's stance on free will by analysing the relation between determinism and the causal role the will plays in bringing about action, I think that Nietzsche turns to a more practical perspective, replacing the traditional dichotomy of free and unfree will with strong and weak will (BGE 21). To establish a positive account of Nietzschean free will, I draw on an analysis of the origins of "freedom" and "the will," and examine how Nietzsche develops the Stoic idea by emphasising the active embrace of suffering and fate as essential to autonomy. Nietzsche's practical understanding of free will as a strong will in the practice of amor fati—one that actively creates its own values amid the constraints of existence challenges the traditional notion and provides a new way to approach the problem, with the potential to liberate the debate from endless conceptual analysis.

A nuanced libertarian approach to free will, focused on perspective

Joel Parthemore

The dominant view among philosophers, cognitive scientists and neuroscientists is that the universe is (strictly) deterministic, with all events fully causally determined by past events. Of those, a majority subscribe to compatibilism: the idea that a substantive form of free will is compatible with determinism, providing one is willing to let go of requiring the capacity to do otherwise, or to have done otherwise.

That “capacity to do otherwise” is exactly what libertarians insist on. That is to say that human beings (and other potential agents with free will) make choices that change the course of events such that they could have chosen otherwise, given no other changes in the physical universe. The universe may not be driven by chance, but it is at the same time non-deterministic.

Libertarians reject the free will of compatibilists as not just very different from what most people understand by free will – thus requiring a different name – but badly motivated and ultimately wrong. The problem with determinism lies in part, they would say, in an overly simplistic view of the nature of causation and, in particular, a naive assumption of a straightforward linear chain whereby a particular set of preconditions mandates a particular set of outcomes: a view that quantum mechanics may be seen to challenge.

I argue for a form of libertarianism that is, so far as I know, original and distinct, as well as (given certain metaphysical assumptions) consistent with observations of the physical universe and human decision making – in part by viewing the causal relationship between subpersonal brain processes and conscious decisions not as linear but circular. Subpersonal events in the brain causally shape conscious decisions *that in turn* causally shape subpersonal events – with no way to determine where the cycle begins. In part, the argument

relies on our limited understanding of the nature of the physical. In part, it relies on the fine line between identity and mere approximation.

To wit, in many if not most contexts in which agents believe themselves to have free will, their actions are sufficiently constrained as to approximate (though, *pace* the standard compatibilist account, never quite equate) to the deterministic model. Through propagation of constraints, previous actions and events constrain present actions to the point where choices, though real, are extremely limited, and only one "choice" may, practically speaking, be possible.

Nevertheless, human beings do have free will in a strong libertarian sense; the problem is only that they mislocate it. They think they have free will (even a great deal of it) in areas where they do not, and fail to see it where they do.

Contemporary libertarian and compatibilist accounts alike compound the problem: compatibilists by equating free will with the ability to act in accordance with one's intentions (even if those intentions are effectively neurally predetermined); at least some libertarians by failing to take note of just how significantly most conscious decisions are constrained, both at the neural level and by social and physical environment; compatibilists and libertarians alike by failing to distinguish between pre-reflective consciousness, where most conscious decisions are arguably made, and fully reflective self-consciousness.

Although my account does allow that, at least in certain circumstances, human beings and like-minded agents have multiple sensorimotor engagements to choose among – in a way that is physically defined but not physically predetermined – nevertheless the heart of free will lies elsewhere, in the capacity to choose one's perspective on oneself and one's surroundings. Both the importance and the power of perspective get overlooked: more than most other conscious mental processes, perspective is, I claim, substantially causally less constrained or predetermined. By selecting among different perspectives on the same circumstances, agents with free will really do change their world -- potentially in profound ways.

The bottom line is that the observable universe only approximates to determinism and, in particular, certain configurations of matter and energy have the capacity to generate a small but appreciable amount of causal force *sui generis*: that is to say, substantially constrained but not fully determined by prior states of the universe. Such entities, commonly recognized as moral agents, can nudge themselves successfully down one choice path or another. Importantly, this view requires nothing "spooky", only letting go of an explanatorily closed system. It requires postulation of nothing non-physical (whatever that would mean), only an acknowledgment that understanding of the physical and all it entails is limited. It does not violate the causal-closure thesis, provided one accepts that a strictly linear view of causality is oversimplification and the physical is more than we can take in.

This account makes no attempt to establish itself as the "correct" view, only as being at least as logically plausible as the compatibilist or incompatibilist-determinist positions – with reasons why one might favour it. It offers opportunity to address a problem with compatibilism that has received insufficient attention: namely, that compatibilists are arguably if inadvertently guilty of a form of the very Cartesian dualism they rail against, by appearing to distinguish between outwardly observable events and "internal" mental events (such as choosing one perspective over another). In a strictly physical universe, mental events must be every bit as much physical as any other events, bound by the same rules. It offers a plausible account of "capacity to do otherwise" free will that does not – unlike Robert Kane's account – rely on quantum indeterminacy. Like Kane's account, it offers a nuanced approach that unapologetically borrows much from the determinist view even while ultimately rejecting it.

Beyond free will: Reconceptualizing responsibility and behavior regulation within determinism

Ting Huang

Keywords: Free Will, Determinism, AI

Abstract

Compatibilists have long sought to resonate a deterministic framework with the idea of free will though redefinitions of freedom. Notions from intentionality, rationality, volitions, control to functional agency etc. have been drafted as “definitions” of freedom. Central to these efforts is the conviction that binds moral agency/responsibility (however limited it might be) to free will. However, it does not have to. This paper seeks to ground our agential choices without invoking metaphysical free will; it proposes that while free will is not compatible with determinism, its social consequence (so-called “moral” responsibility) is. In other words, it argues against free will and retains “moral responsibility” under a different name - as we are not “free”, responsibility cannot be called “moral” anymore - rather it will be reframed as consequence for action (CFA) in this paper, a pragmatic tool for influencing behavior rather than a metaphysical judgment of our status.

To be more precise, human, as well as other beings, will be seen as an input-output system - we act in accordance with the causal outcome of internal processes responding to inputs. Within this framework, CFA operates as an external input - a regulatory mechanism designed not for retributive punishment but for behavior modulation. Under this lens, the social contract does not rely on metaphysical freedom but on the pragmatic need for accountability to maintain order for social living.

Along this line of thought, discussions of free will in AI will not be meaningful, as free will is non-existent period - rather when and under what criteria do we attribute CFA to systems. For navigation, this paper’s will say yes to human consciousness, yes to human agency (as in ability to make rational choices) but no to human freedom (the same input will generate the same output). And it will say not sure to artificial consciousness (partly due to Ned Block’s influence but answer to this question does not affect what’s being discussed here), yes to artificial agency (so both artificial and biological brains can make rational decisions based on input) and no to artificial freedom (again, same input, same output).

The question of whether CFA should be attributed to AI introduces another vital component that ensures the functionality of CFA: the survival instinct. For CFA to be effective, the entity it applies to must be motivated to avoid the consequences, for biological systems it is an inclination rooted in the biological imperative to survive. The current models of Artificial Intelligence does not possess this, thus its meaningless to attribute CFA to the current models at least - if it does not “care” that you pull its power plug, CFA is not gonna be able to modulate its behaviour.

The question of whether CFA should be attributed to AI introduces another vital component: the survival instinct. For CFA to be effective, the entity it applies to must have a reason to avoid the consequences. For us it is an inclination rooted in the biological imperative to survive. Current AI models lack such intrinsic motivation, thus it will be meaningless to attribute to the current models - if they do not “care” if you cut their power source or threaten their existence, CFA is not gonna regulate its behaviour.