Special Issue, "What AI Can Learn from Biology"

Vol. 8, No. 1–2 (2025)
ISSN: 2532-5876

DOI: 10.13133/2532-5876/18906

Could Artificial Intelligence (AI) Become a Responsible Agent: Artificial Agency (AA)?

Raymond Noble[a] *and Denis Noble*[b*]

[a] *Honorary Senior Lecturer, Institute for Women's Health, University College London, United Kingdom*

[a] *Emeritus Professor, Department of Physiology, Anatomy and Genetics, University of Oxford, United Kingdom*

***Corresponding author:** Denis Noble, Email: denis.noble@dpag.ox.ac.uk

Abstract

Responding to concerns that superintelligent AI could escape human control, this paper argues that the true existential question is not intelligence but agency, and that artificial intelligence as currently conceived poses no threat of responsible agency. Intelligence can be fully artificial and beneficial (books, databases, algorithms) without ever bearing responsibility. Responsibility belongs exclusively to agents, specifically biological agents. Biological agency requires causal independence, intentionality, creativity, and above all the active harnessing of stochasticity to generate novel, goal-directed behavior that is neither predetermined nor merely random. Organisms achieve this at every level—from ion channels and immune-system hypermutation to neural decision-making and social anticipation—by constraining chance rather than eliminating it. Choice in living systems resembles poker rather than chess: iterative, intuitive, socially embedded, and inherently unpredictable even in principle. Algorithmic systems, even those incorporating randomness, cannot replicate this multi-level process. Creating genuine artificial agency would demand reproducing biology's constrained use of stochasticity across scales. Only then could a machine become a responsible (or irresponsible) agent. If achieved, the distinction between living and artificial would collapse, raising profound ethical questions. Until then, the risk lies not in AI itself but in failing to regulate research that might inadvertently cross this threshold.

Keywords: biological agents, stochasticity, responsibility, intentionality

Citation: Noble, R, & Noble, D 2025, "Could Artificial Intelligence (AI) Become a Responsible Agent: Artificial Agency (AA)?", *Organisms: Journal of Biological Sciences*, vol. 8, no. 1–2, pp. 27–31. DOI: 10.13133/2532-5876/18906

Ali Hossaini's essay raises a question that ought to concern humanity very deeply indeed: could intelligent machines challenge humanity's place on Earth? He is right to question how we detect and regulate the emergence of agency, and agency should be put on the agenda. This is because the threat is not from intelligence as such. Humanity faces no real threat from 'artificial' intelligence. On the contrary, people have benefited enormously from the 'artificial' ways of storing ordered facts and intelligence in books for thousands of years, and in other databases more recently. We have used those tools to our great benefit. Moreover, it is clear where the responsibility lies for the production of the tools. They are other humans, those who wrote the books, and those who created the databases. There are ethical and legal reasons why it is sometimes very important to know who those agents are. It is agents who carry responsibility, not dead pieces of paper with ordered ink particles, nor the bits of electronic machinery that can harbor databases. If facts are wrong or misleading, or machinery does not work properly, we know who to blame.

They are to blame precisely because they are agents.

As Hossaini's essay also says, there is even a disconnect between intelligence and agency. Desire is often in defiance of logic. So, what is agency in organisms?

In this response, we outline what is required to be an agent and why it may be difficult for machines to be made that could have agency. If that could be done it would raise ethical issues on how we treat and interact with them.

1. What is Agency?

Agents can choose and anticipate the choices of other agents. Furthermore, they can do so creatively, and not simply by following a predetermined algorithm. To quote from one of our recent articles (Noble & Noble 2018):

> An agent acts, it does not just react in the way, for example, in which a billiard ball is caused by another ball to move. There are many levels of agency (Kenny 1992, pp. 32–40). Organisms are agents to the extent that they can interact socially with other organisms to choose particular forms of behavior in response to environmental challenges. Agency requires causal independence (Farnsworth 2018). It also requires intentionality, i.e., the sense of purpose, in order to be causally effective as a driving force (Liljenstrom 2018).

Agency also involves iterative forms of anticipation, as we will show later in this article. Determinate algorithms or sets of algorithms alone cannot do this.

A purely stochastic system might be defined as one in which all states are equally possible. Thus, all the possible combinations of two unbiased dice would occur by chance equally frequently. However, variations in biological systems are constrained and utilized to generate particular outcomes that are not as equally probable as all other possible outcomes. Precisely this gives the system the potential to be creative. The system uses chance, but the outcome is not pure chance. It is goal-directed. This is what we mean by agency. In the same article we outlined an empirically testable theory of choice based on the active harnessing of stochasticity:

> For an empirically testable theory of choice to be possible, we need to know at which stages in the process experimental interventions could test its validity. At first sight, that may seem impossible. How can we specify a process that is necessarily *unpredictable* but which can be given an at least apparently *rational* justification once it has happened? Our previous work provides a clue to that problem (Noble & Noble 2017). We analyzed agency by comparing it to the purposive behavior of the immune system. The immune system solves what we can best characterize as a template puzzle: given a new invader with an unknown chemical profile (shape of template), what is the best way to find the key (an anti-template, i.e., the antibody) to lock onto and neutralize the invader? The answer in the case of the immune system is one of the most remarkable forms of the harnessing of stochasticity. In response to the new environmental challenge, a feedback loop activates a massive increase in mutation rate in a highly targeted region of the immunoglobulin DNA sequence (Odegard & Schatz 2006). The process of choice in organisms can be viewed as analogous to the immune system.

Choice and anticipation require the harnessing of stochasticity. An important part of our argument is that the use of stochasticity in biology has been

misunderstood. The standard theory of evolution (neo-Darwinism), for example, treats random variations in DNA as simply the origin of new DNA variants, with absolutely no control by organisms themselves. They are viewed as the passive recipients of such variation. Choice between the variants is then attributed to the process of natural selection.

By contrast, we argue that organisms actively harness stochasticity in order to generate novelty in their behavior from which they can then select to best meet the challenges they face (Noble 2017).

Challenges facing organisms can be viewed as a puzzle analogous to the form of a template for which a match is needed. The challenge might be a routine one, in which case what we *normally* characterize as a reflex, or predetermined response, may be adequate. It might be considered that such a response would *not* involve a choice although, even so, biological systems often act to allow this to occur. Any artificial system would need to replicate such choices, and it would also need to replicate the kind of choice involved when no automatic reflex response is possible. The challenge facing the organism then is what could fit the puzzle template?

We speculate that stochasticity is harnessed throughout the processes used by the organism to achieve this.

For cognitive problems in organisms with highly developed nervous systems, these will be primarily neural. Neural processes are extensively stochastic at all functional levels, from the opening and closing of ion channels via action potential generation, spontaneously or through synaptic transmission in neuronal networks, up to cognitive functions, including decision-making (Hille 1992; Heisenberg 2009; Tchaptchet, Jin, & Braun 2015; Brembs & Heisenberg 2018; Braun 2018). Furthermore, harnessing stochasticity underpins the function of all living cells. It generates the membrane potential necessary for the electrochemical function in all cells.

A further speculation is that, once the harnessing of stochasticity has thrown up possible novelty, the organism controls the next stage, which is to compare the novel options with the problem template to determine what fits. 'Template' and 'fit' here are used metaphorically, in much the same sense in which a logical answer can be said to 'fit' (that is to say, answer to) the problem posed by a question. This is the essential choice process, needing a comparator.

Our theory is an idealized process, but it clearly helps to explain an apparent paradox regarding the predictability or otherwise of what we call a free choice. The logic lies in the fit between the problem template and the solution template. But the stochastic stage of the process ensures that the choice may be unpredictable since we cannot predict what stochasticity will throw up. So, free choice can be both rational and novel.

Stochasticity is harnessed throughout the process. This is characteristic of biological systems. While not impossible, it may be difficult to construct AI systems that can replicate this. If and when AI could mimic biology then it would raise a fundamental problem: would this system be living?

If so, the distinction between artificial and natural would disappear.

'Rational' here does not necessarily mean the most logical choice. As Laurie Santos and Alexandra Rosati write, "we now know that human choice is often not as rational as one might expect" (Santos & Rosati 2015). This is necessarily true since, within the context of the choice process, there is obviously no guarantee that a stochastic process will throw up a fully rational solution. Partial success is what would be expected most of the time. The same is true of the immune system. All it needs to do is to come up with a 'good enough' template match. It does not have to be the perfect match. If a key fits the lock, it does not really matter whether it is an exact fit.

How then do humans come to feel that their 'imperfect' but 'effective' choices really are theirs? After all, most of the time we can give a 'good enough' explanation (the rationale) for a choice, however partial the 'fit' may seem to be to the problem. A possible solution to that problem could be what Santos and Rosati call the endowment effect. We privilege retaining what we already own. By 'rational' here we do not mean 'the most intelligent response'. It means only that the decision was rational to the agent in the sense that the agent owns the response he chose to make.

2. The Logic of Social Interactions

All organisms utilize stochasticity in creative responses to change. This is achieved in a continuous process of iteration and re-iteration. They do this at many different levels from the molecular (immune system cells activating hypermutation) to the level

of whole organisms (bacteria using those molecular processes to evolve their immunity to antibiotics) through to the social levels. It is at a social level that we can talk of reason in terms of social motivation.

Consider why Jack went up the hill. He may have done so not only to fetch a pail of water, but because he wanted to be with Jill, with whom he had fallen in love. If we tried to model this mathematically, it would be exceedingly difficult because there are so many initial and boundary conditions. Much of Jack's behavior is in anticipation of Jill's; and Jill's of Jack's; and even what they believe others might think of them. It is at the social level that shared concepts of right and wrong might influence choices. An agent at such a level might anticipate that another may act in a way that might be considered wrong, and in turn predicate choices on such possibilities. There is a continuous process of adaptability in the choices made; a continual process of assessment of whether or not the right choice has been made. Furthermore, the 'right' choice may not be made; we make 'mistakes'; we take the 'wrong' turning; and this also is part of our intellectual endeavor. We mold our decisions in the process of carrying them out. We try things out, and sometimes make a choice by a mental toss of a coin. We may stick with a choice simply to see what the outcome will be.

Agency in organisms is therefore more like a game of poker than a game of chess. In chess at least the type of move is restricted and known; in living organisms this is not so readily the case. A pawn may be moved in a very restricted number of ways; a bishop can move diagonally, but is nonetheless restricted, although it might not be clear how far it might be moved. There are nonetheless 'rules' of the game. But what if the game has no such rules, or that the rules are indeterminate. In particular, in the light of what we have written above, they may be indeterminate, because 'chance' or stochastic processes are utilized in deciding a move. An algorithm could work only in as far as it gets us to the point of saying, "if X then spin the wheel of chance". A buffalo may anticipate the mood of the lion; it may also anticipate which way the lion may turn; the lion also anticipates the anticipation of the buffalo; to varying degrees, each is spinning a wheel. Each is 'reading' the other, but almost always with uncertainty.

Anticipating is not a simple calculation, it is intuitive; it is based on the assumption that something is not calculable. We cannot measure the strength of Jack's love for Jill; we know it influences his behavior, but we do not know precisely its strength in any given moment or event. Yet, it is a factor in our deliberation of his likely responses. Desire, lust, anger, hate, pain, and so much more influence his actions, and these ebb and flow, often in unpredictable ways. If a driver of a car reaches a junction at which he is momentarily blinded by the sun, all such factors and more might influence his decision. We might understand his character traits, what he is likely to do, but we are unsure in any given incidence. Living organisms work with uncertainty. John always obeys the 'law' and never knowingly jumps a red light; Peter sometimes will, but not always; and even John might if after time he concludes that the traffic light is no longer working. When will a 'rule' be broken? Life anticipates it might be. If we did create artificial agency, then we would have to live with its uncertainty. If we made AI that merely obeys our will or is entirely predictable then it cannot have agency. It is simply a tool. That would be true even of an AI system that merely includes stochasticity without the harnessing process. Such a stochastic algorithm would have been placed there by humans, not actively developed by the organism itself.

This point is related to part of the basis of Donald MacKay's argument in 1960 for the logical indeterminacy of a free choice (MacKay 1960). To quote MacKay:

> For us as agents, any purported prediction of our normal choices as 'certain' is strictly *incredible*, and the key evidence for it *unformulable*. It is not that the evidence is unknown to us; in the nature of the case, no evidence-for-us at that point exists. To us, our choice is logically indeterminate, until we make it. For us, choosing is not something to be observed or predicted, but to be done. (MacKay's own emphases)

MacKay also writes:

> In retrospect, of course, the agent can join the onlookers (e.g. in witnessing a moving film of his own brain processes) and share in their 'outside' view of his physical past as 'determined'. Past and future have an asymmetric logic for an agent.

We mostly agree with MacKay on both of these conclusions, but it is important to note that MacKay does not include the importance of harnessing stochasticity in the formation of a free choice. On the contrary, he refers to the agent's physical past as 'determined'. That is an important omission since including the harnessing of stochasticity means that any 're-running' of his imagined brain film would not necessarily lead to the same outcome. In our view of the nature of a free choice, there can be many 'rational free choice' fits to same challenge. So the agent could indeed join the onlookers in watching the film of what actually occurred, but he would still be able to assert that his action was not predetermined. Our social being also allows us to learn by mistakes. It is part of our intelligence. Our intelligence is cultural and transgenerational, and it allows a spinning of the wheel in ways beyond simply the organism. Our social being buffers us from mistakes in the choices we make. It allows protection while we take time to deliberate, to consider alternative courses of action. It allows us to learn from the mistakes or successes of the past. It also allows us to take a collective decision, and to argue about it. AI researchers have recognized this and have made progress in seeking to replicate it (Arulkumaran *et al.* 2017). It allows us to spin the wheel politically. All this is part of our being as intelligent agents, and we may harness the power of AI to test new ideas about our world. Our complex mathematical models of living systems are impossible to understand without the calculations available in modern computers. The use of AI is part of our spinning the wheel.

Conclusions

The functional harnessing of stochasticity is essential to life as we know it. It occurs even in the prokaryotes, bacteria and our own ancestors the archaea. It is essential to agency, for otherwise there would be no creativity in the behavioral repertoire of living organisms.

In order therefore to reconstruct agency, AI research will need to find ways of incorporating the harnessing of stochasticity, as organisms do and have done for billions of years. To achieve this, it will not be sufficient simply to add stochasticity to otherwise deterministic algorithms. The functional multi-level *harnessing* process must also be reproduced.

Who knows, we might then even be able to fall in love with a future AI robot. Perhaps we would no longer call it a robot.

Meanwhile, the threat should not be taken lightly. It is a real threat to humanity and it requires careful regulation. We already know the price of not regulating the free exploitation of AI. We cannot afford to wait until IT research actually succeeds in producing non-human agency – if indeed that is possible.

References

Arulkumaran, K *et al.* 2017 "Deep reinforcement learning: A brief survey", *IEEE Signal Processing Magazine*, vol. 34, pp. 26–38.

Braun, H 2018 "Der Zufall in der Neurobiologie - von Ionenkanälen zur Frage des freien Willens". In: Herkenrath, U (ed.), *Zufall in der belebten Natur* (Hennef: Verlag Roman Kovar), pp. 109–137.

Brembs, B, & Heisenberg, M 2018 "Der Zufall als kreatives Element in Gehirn und Verhalten". In: Herkenrath, U (ed.), *Zufall in der belebten Natur* (Hennef: Verlag Roman Kovar), pp. 80–94.

Burns, B 1968 *The Uncertain Nervous System*. London: Arnold.

Farnsworth, K 2018 "How organisms gained causal independence and how it might be quantified", *Biology*, vol. 7, no. 3, Article 38.

Heisenberg, M 2009 "Is free will an illusion?", *Nature*, vol. 459, pp. 164–165.

Hille, B 1992 *Ionic Channels of Excitable Membranes*. Sunderland, MA: Sinauer Associates Inc.

Kenny, A 1992 *The Metaphysics of Mind*. Oxford: Oxford University Press.

Liljenstrom, H 2018 "Intentionality as a driving force", *Journal of Consciousness Studies*, vol. 25, no. 1–2, pp. 206–229.

MacKay, D 1960 "On the logical indeterminacy of a free choice", *Mind*, vol. 69, no. 273, pp. 31–40.

Noble, D 2017 "Evolution viewed from physics, physiology and medicine", *Interface Focus*, vol. 7, no. 5.

Noble, R, & Noble, D 2017 "Was the watchmaker blind? Or was she one-eyed?", *Biology*, vol. 6, no. 4, Article 47.

Noble R, & Noble, D 2018 "Harnessing stochasticity: How do organisms make choices?", *Chaos*, vol. 28, no. 10.

Odegard, V, & Schatz, D 2006 "Targeting of somatic hypermutation", *Nature Reviews Immunology*, vol. 6, no. 8, pp. 573–583.

Santos, L, & Rosati, A 2015 "The evolutionary roots of human decision making", *Annual Review of Psychology*, vol. 66, pp. 321–47.

Tchaptchet, A, Jin, W, & Braun, H 2015 "Diversity and noise in neurodynamics across different functional levels". In: Wang, R, & Pan, X (eds.), Advances in Cognitive Neurodynamics. Singapore: Springer, p. 681–687.