HUMAN SUPERINTELLIGENCE:

How you can develop it using recursive self-improvement

Extract from Chapter 22: The Meaning and Purpose of Life and Suffering

He began to talk to me about the need for human beings to believe in something much larger than themselves that would continue to exist long after they died. This could give them meaning and purpose in their life, he suggested. My immediate reaction to this was that the evolutionary worldview provided me with abundant meaning and purpose—probably more than was experienced by most others on the planet. It provided me with a role and a purpose in a much larger scheme of things that will outlive everyone I know.

His response was yes, but did I not accept that there is an ineradicable mystery at the heart of human existence that even the evolutionary worldview cannot explain? I responded that yes, I accept this, but for me, it is a very circumscribed mystery—once the big bang occurred, there was no mystery as to how the universe unfolded: known physical laws and material processes eventually gave rise to life; this life increased progressively in hierarchical complexity and evolvability; evolution eventually gave rise to a cognitive capacity that could comprehend the processes that produced this evolutionary trajectory; ultimately this enabled a shift to intentional evolution; and so on, and so on.

But I had long recognised that none of this explains why there is a universe in the first place. I had accepted that this mystery was impenetrable. Evidence did not exist that enabled science to peer beyond the veil and to develop scientific explanations of why there is something rather than nothing.

I knew that it was possible to construct an infinite number of hypotheses that each could explain logically why the universe exists and that cannot be contradicted by any known facts. For example, all sorts of gods with all sorts of powers could be hypothesised, as could multitudes of possible material causes. However, there is no evidence that could be used to falsify any of these hypotheses or to narrow them down to a single hypothesis.

Furthermore, if some means were found to enable science to develop an understanding of what generated the Big Bang, this would push back the horizon of our knowledge somewhat, but there would still be a horizon beyond which we could not know anything with certainty.

Consequently, I agreed with him that there was an ineradicable mystery at the heart of human existence and, in fact, at the heart of any existence at all. However, at first, I did not see that this had any significant implications.

But as his questions wormed their way into my mind, they connected with other realizations and intuitions. In particular, they resonated with the messages I received from the voices I heard during my mushroom experience: there are larger-scale processes at much higher levels than ours, and I do not know much about what is really going on here.

Previously, I had rejected spending much time thinking about what might have produced our universe in the first place. I believed that doing so was a dead end. It was impossible to make any progress in sorting through the endless possibilities.

However, I began to realize that I had been using ineffective methods to address these issues. I had been proceeding as if my primary goal was to develop an objective and scientific explanation of the causes of the universe. But now I began to understand that I should have seen this approach as just one of several possible means for discovering what I really wanted to know, not just an end in itself.

In fact, my central interest was whether the nature of these causal processes might have implications for us, here and now. My main goal was to discover whether an understanding of the processes could answer questions like: Was our universe established for a particular purpose? What role, if any, do we have in fulfilling this purpose? Will it make any difference to us whether we consciously set out to do what we can to assist the achievement of this purpose?

Of course, if a scientific approach could explain in detail what caused our universe to emerge, it would greatly facilitate the answering of these and other questions about the sense and significance of our existence. Once science understands these causes sufficiently, its discoveries can be used to work out whether there are any implications for the way we live our lives.

However, we can probably never know with scientific certainty the precise nature of the processes that produced our universe. We are unlikely ever to be certain about whether these unknown processes have major implications for our lives.

In the absence of a scientific explanation, the possibilities are endless. Depending on the nature of the possibilities, what we do during our lives might make no difference. For example, there may be no particular way in which we can act that will pay off for us in the longer term, e.g. by enabling us to survive after death. But neither can we rule out possibilities that would provide such payoffs. We can imagine numerous plausible possibilities that have the potential to provide meaning and purpose for at least some forms of human existence.

Nonetheless, in the face of this uncertainty, are there rational methods we can use to decide how we should live our lives now to maximize our longer-term interests? Increasingly, I realized that this was the most relevant question to ask about the fundamental nature of our reality.

Even if there is radical uncertainty about why there is something rather than nothing, is it still possible to make rational decisions about how we should act now? Is there a method for making choices that does not rely on certain knowledge, but that, for example, maximizes our chances of surviving in the long term, including after our bodies die?

Of course, humans make decisions regularly in the face of uncertainty. Techniques have been developed and tested that enable them to do so rationally and effectively, even in the face of radical uncertainty. Radical uncertainty exists when it is impossible to predict accurately the future consequences of possible actions, or even to assign probabilities to the possible outcomes. Humans face radical uncertainty about the reasons for the existence of our universe and their implications for our lives.

These decision-making methods can be particularly effective when at least some of the possible outcomes of decisions can produce significant benefits for the decision-maker. As we shall see, this is the case even though there may be many more possible outcomes that do not impact on the decision-maker at all.

An infinite number of possibilities exist that are consistent with all available evidence that could explain the existence of our universe. Some of these possibilities have the potential to make considerable sense of our universe in general and of human existence in particular. Significantly, these meaningful possibilities cannot be ruled out. They are actual possibilities that seem plausible and that are not falsified by any known evidence.

I began to realize that it is possible, for example, that the larger-scale processes that are responsible for the existence of our universe may have set it up for the purpose of generating higher intelligences. The path that our universe provides for the development of these capacities might include accepting suffering and learning as much as possible from enduring it. If this were the case, and it cannot be ruled out, the option of turning off life support might not be as rationally attractive as I thought previously. It may involve closing the door to further possibilities that are valued and even rewarded by the processes that established this universe.

This perspective might be seen to differ from my original evolutionary worldview in only minor ways. However, it made a huge difference to how I would come to view my life and whether I should persevere with it even in the face of great on-going suffering.

I will now set out to identify in detail the rational decision-making methods that are capable of identifying optimal strategies despite radical uncertainty. I will then apply them to these fundamental existential issues.

But such strategies are only valid if there are plausible possibilities that cannot be ruled out and that have longer-term implications for what humans do here and now in the world. I will begin by demonstrating the existence of a number of such significant possibilities.

Among possible explanations for why our universe exists are those that suggest it was brought into existence by a being(s) that exist outside our universe and have the capacity to create universes. Their methods of creation might include producing simulations.

Why would such a powerful being(s) create our universe? It is possible to divide the possibilities into several relevant, overlapping categories. First, there are an infinite number of possible explanations that make no sense in the context of human intentions and goals. Furthermore, an infinite number of possibilities exist that have no implications for how we live our lives. No matter what we do in such a universe, our actions would have no consequences for us beyond our temporary existence. There are also an infinite number and variety of 'God hypotheses'.

However, there is a class of somewhat plausible possibilities in which the being(s) who brought our universe into existence have done so for their own particular purposes. When they established our universe, they set it up in such a way that it performs particular functions that serve their goals. But of course, it might not have been set up at all to serve human goals. No matter how we live our lives in such a universe, it might make no difference to whether we can, for example, survive death.

For example, our universe might be a simulation established by an advanced civilization that exists outside our universe. In this hypothetical example, they have established the simulation for the purpose of evaluating the consequences of alternative ways of organising their societies. It may be that because of computational irreducibility, even such advanced intelligences cannot work out the consequences for their own civilization of these alternative strategies. Consequently, they need to simulate various possibilities if they are to identify the best strategy.

This advanced civilization may be completely unconcerned that their simulated universes will produce extensive misery and suffering for the conscious beings that will emerge in the simulations.

But there is a class of possibilities that is more relevant and interesting, at least to humans. In these cases, the way we live our lives does have consequences for the possibility of some kind of existence that extends beyond our bodily deaths.

I will briefly outline three classes of this kind of possibility that cannot be ruled out.

The first class of possibilities is the creation by an advanced civilization of simulations that are designed specifically to provide intelligences with environments and experiences that enhance their psychological growth and development.

The plausibility of these possibilities is illustrated by the fact that humanity is already beginning to head in this general direction.

Human therapeutic psychology has often used visualization, imagination, and environmental manipulation to heal individuals from trauma and to build greater psychological resilience and functioning. In recent years, simulations and virtual realities have also begun to be used to deliver these kinds of therapies, albeit in a limited way. A very simple example is the use of virtual environments to extinguish phobias associated with airplane travel. The virtual environment gradually introduces the individual to the experience of being a passenger on a plane. It exposes them progressively to the negative feelings triggered by their phobia, but in an environment that keeps them calm and relaxed. Once they get used to a particular level of exposure, it is increased further, eventually extinguishing the phobia.

However, the appropriate use of complex and realistic virtual environments is likely to take this to an entirely new level. In principle, individuals could be temporarily embedded in whatever social and other environmental circumstances will produce the desired psychological outcomes.

An appropriate set of experiences could significantly change an individual's conditioning. Or it could operate at a meta-level, propelling the development of a capacity to be self-evolving. This would almost certainly have to include experiences that are traumatic for the individual, thereby motivating them to leave the comfort of their existing level of psychological development and to do the hard work on themselves that is needed to move them vertically to the next level.

Of course, in order to have maximum effect when an individual is embedded in such a virtual reality, it may be desirable to induce the individual into a state in which they believe that their virtual experiences are real. When this is achieved, the participating individual would experience the virtual therapeutic experience as being indistinguishable from real life.

This kind of technology might become very significant if humans achieve a state of abundance. Many dream that eventually, human technological capabilities, perhaps aided by AI, will be able to effortlessly meet all human needs and wants. If such a society were ever achieved, children might need to spend considerable time in a virtual reality that provides them with the shocks, traumas, and other complex challenges that are essential for successful human psychological development.

Our universe could be such a simulation created by an advanced civilization.

The second class of possibilities that I will consider is the creation by an advanced civilization of simulations that are designed to develop higher intelligences in a way that is safe for the civilization.

Again, human AI developers are already beginning to consider the desirability of such an approach. This is because AI might emerge eventually that has the power, motivation, and intelligence to harm its creators. Some believe that it may even destroy human civilization.

A solution might be to confine the training and development of AI to simulated environments that provide no opportunity for the AI to interact with the creators or their universe. These simulated 'sandbox' universes would be specifically structured so that they facilitate the development of AI of higher intelligence. Once particular intelligences reach the desired level, they could be harvested by the advanced civilization and used for their purposes.

However, harvesting would proceed only once the AI was assessed as being safe from the perspective of the advanced civilization. Additional security could be provided by embedding the simulation in a nested hierarchy of simulated environments. Escape from lower-level simulations would not threaten the civilization.

From the perspective of an advanced civilization, we might be AI that is evolving and developing in a sandbox universe that it has simulated.

The third class of possibilities arises because it may be impossible to produce highly intelligent AI by a combination of engineering and training. The only viable way to produce such AI might be to set up a simulation of an entire evolutionary process. This process would begin with the emergence of a universe. Eventually, life would emerge within the simulation and evolve in the direction of increasing complexity and evolvability. If it were set up appropriately, such a simulation would eventually produce high-level intelligences.

This class of possibilities cannot yet be ruled out. This is despite the fact that many AI researchers believe strongly that it is possible to engineer and train Artificial General Intelligences (AGI). However, none have done so, or have come even close. Human attempts to produce AGI or even to understand how it might function have not yet gone beyond wishful thinking.

At present, we know of only one particular strategy that can reliably produce human-level intelligence in an entity that begins far below this level. The strategy's starting point is a newborn baby. Typically, the baby cannot speak, think abstractly, coordinate its bodily movements, or undertake the other myriad functions that characterise human-level capacities.

Thirty years later, and with a bit of luck, the typical baby will have grown into an adult with human-level intelligence. In the early years, the physical maturation of the brain and other bodily systems will have contributed to this development. But mostly, the intelligence of the growing human will develop due to learning and associated processes that help to install the relevant capabilities.

What kinds of experiences, interactions, and training does a growing human need if its intelligence is to develop appropriately? Could an AI that begins with baby-level intelligence be subjected to similar experiences, interactions, and training, thereby achieving comparable improvements in intelligence? Furthermore, are there alternative approaches that could be taken with AI that could short-circuit our current methods of producing human-level intelligence? For example, to what extent could AI be programmed directly with capacities that humans have to learn?

The current state of human knowledge about how to develop human-level intelligence in humans is very limited. It happens right in front of us, but we actually know very little about how to do it in humans, let alone in AI. As I have outlined, even I made one or two mistakes while raising my daughters.

However, it is possible to get a general idea of what is necessary to produce human-level intelligence in humans. This can be achieved by identifying in broad terms the kinds of learning experiences that children are subjected to as they develop. We can get a sense of the learning experiences that are most important by noting those that are essential for the successful development of the child.

It is well known that the foundations of our cognitive and social-emotional development are established in the first year or so of life through long sequences of interactions with our mothers or other primary caregivers. This then broadens into the complex processes of socialization that generally involve other family members and then wider communities. Again, this comprises innumerable interactions and learning experiences. We have little idea about which particular interactions are critical for our successful development. But we know, in general, that these kinds of complex interactions and learning experiences are essential if we are to continue to grow and develop successfully.

In a modern complex society, this is followed by immersion in pre-school and kindergarten, and then 12 long years of schooling. Without something like this, we will not develop the cognitive and social/emotional knowledge and capacities to get well-paid work in a modern economic system.

But formal education is not the only critical factor during these years. Play and other complex interactions with peers that occur before and during school are also significant in developing all kinds of social/emotional and cognitive skills.

After schooling comes university or work and entanglement in a wider community. Again, myriads of interactions and learning experiences occur. It is not possible to identify the specific sequences and networks of experiences that eventually prove essential for enabling a particular career.

Whatever career we eventually find ourselves suited to, it may enable us to make a significant contribution to the complex society in which we live and work. However, all the intervening steps that took us there could not have been planned in advance. Many skills and capacities that prove significant in our lives are acquired due to chance meetings with individuals who grew up in completely different environments or even in different countries and cultures.

Failures, traumas, abusive relationships (including within the family), and mental illness may, in retrospect, have been critically important for driving our acquisition of skills and knowledge that eventually prove essential for our development. Would we develop fully as intelligent human beings without encountering challenges that cause us to, for example, engage in self-reflection; experience failed friendships; fall in and out of love; learn from bitter experiences that our behaviour when younger was inappropriate; realize many years after this that the revised behaviours we adopted were also inappropriate; move through Piaget's levels of cognitive development; learn how to meditate to accelerate this developmental process; realize that an unexamined life is not worth living; develop psychological defences and then modify them and eventually drop them; and so on, and so on?

How could we possibly provide AI with these experiences that appear essential for developing intelligence capable of achieving complex goals in our world? We would not know where to start.

At present, humans have little knowledge about which kinds of experiences might be essential to develop our cognitive and social-emotional capacities. As this book outlines in relation to my life, the importance of these experiences might not be at all obvious as they occur. As Soren Kierkegaard said, "Life can only be understood backwards, but it must be lived forwards."

In some tribal societies, it was said that it takes a village to raise a child. Now, it tends to take at least a modern nation-state. Increasingly, it is taking an international system of markets and economic activities. Eventually, it will take a living global entity.

Is this what we must do to produce human-level AI and beyond? Like us, will AI have to grow and develop in a diverse and complex society if it is ever to acquire human-level intelligence?

Do we have to begin with AI that is roughly at the same level as a human baby, and then embed it in an environment that will deliver it the myriads of learning experiences and interactions in the 'correct' sequences that eventually transform human babies into fully-functioning adults?

If so, it seems impossible to do this by designing each of these experiences and then delivering them to the AI in a training environment. Attempts to do so would be likely to be continually beset by computational irreducibilities.

It should also be obvious that this cannot be achieved by providing the AI with all human declarative knowledge that is available on the Internet and elsewhere. This will go nowhere near providing the developing AI with the myriads of complex social and physical interactions that humans experience as they develop.

Instead, it seems that the only way to proceed successfully would be to embed an agentic and goal-directed AI into a complex social environment that self-organises the necessary experiences, in interaction with the AI.

Given our current knowledge and experience, the only feasible way of doing this would be to raise the AI 'baby' as if it were a human baby. At present, the only way we could have any hope of ensuring that a developing, agentic AI will have learning experiences similar to those that can transform a human baby into a functioning adult would be to immerse it in the only environment that can currently provide such opportunities—a human society.

A similar approach was taken to test whether primates could attain human-level capacities if they were provided with appropriate learning opportunities. Primate babies were brought up in human families. However, they did not progress far.

But even if such a strategy were feasible, before we can begin to implement it, we encounter an even bigger challenge. We may not even be able to get to the starting point that I have assumed in the discussion so far. We have little idea about how to design and engineer AI that is as capable as a newly born human baby and, more importantly, that has the potential of a human baby.

This is a major obstacle. The only process that we know of that can produce anything like a human baby from scratch is the entire evolutionary process itself.

The reasons why this challenge may never be overcome are similar to the reasons why it is unlikely that we can ever produce human-level AI directly by engineering and training: much of the knowledge embodied in a human baby is procedural; it has been discovered by billions of years of evolutionary trial-and-error; computational irreducibilities make it impossible to build models that would enable us to understand and manipulate the relevant processes; as a consequence, we have no idea how to emulate what the evolutionary process has achieved or even to identify which particular steps were critical in moving towards baby-level intelligence accompanied by the potential to develop human-level intelligence; and so on.

However, if we start a universe with a Big Bang that is appropriately fine-tuned, it seems likely that it will eventually produce life, and the evolution of this life will have a trajectory. As we have discussed, the details will differ, but eventually, the trajectory seems likely to produce organisms that have human-level intelligence.

If humanity develops the capacity to simulate the emergence and evolution of such a universe, it seems likely that eventually, we would be able to harvest intelligences at the human level and beyond that develop within the universe.

I have only sketched the relevant arguments and considerations here. But it may be the case that the only way in which an advanced civilization could produce human-level AI and beyond is by simulating the formation and evolution of a universe. Such a simulation would have to be set up with initial conditions that ensure that the evolving universe is both life-friendly and intelligence-friendly and that produces an evolutionary trajectory that heads in the direction of generating life and intelligence of increasing complexity and evolvability.

Of course, if humans decide to proceed down such a path, it would highlight the possibility that our universe and our lives are the product of such a simulation. The argument made by philosopher Nick Bostrom about the probability that we are actually living in such a simulation applies equally here.[1] If producing such a simulation is something that civilizations are likely to do once they become sufficiently advanced, we are probably in such a simulation. The probability that any given civilization is the original one that gave rise to a long sequence of simulations is very small. It is much more likely that it is a simulated one.

But this brings me back to the central issue that propelled my thinking about these issues. To what extent can the existence of these classes of possibilities help address the ineradicable mystery that exists at the heart of human existence? They are just possibilities, incapable of eradicating the mystery completely. No one can explain with certainty why something exists rather than nothing. Why have I gone down this path knowing that it is incapable of producing certainty about the big existential question that we all face?

At best, as we have seen, these possibilities represent hypotheses that cannot be ruled out on the basis of current evidence. However, there are an infinite number of hypotheses that can explain current circumstances and are consistent with all known facts (these include an infinite number that rely on various gods and supernatural beings as well as materialistic theories). Furthermore, most of these possibilities fail to get beyond the infinite regress that arises when we ask the questions: Who or what created those being(s) that are hypothesised to have created our universe, and who or what created them, and so on, and so on, indefinitely?

However, as I suggested earlier and will now demonstrate in more detail, possibilities that cannot be established or ruled out by current evidence can provide good reasons for a rational agent to act in particular ways. This is the case even when there is no evidential basis to establish any particular possibility, or even to assign probabilities to any possibilities. Radical uncertainty of this kind is a common challenge for humans and other agents, and it is not the end of the story.

An example of such a challenge is when an agent is faced with many hypotheses that make predictions about future events. In this example, no evidence establishes any of these hypotheses or even enables probabilities to be assigned. However, consider a hypothesis that makes specific predictions about a particular future event. The predictions are such that if the agent uses the predictions to decide what actions it will take, and if the hypothesis proves to be accurate, the agent will reap considerable benefits.

A specific example is the hypothesis that our universe is a simulation created by an advanced civilization that exists outside our universe. The hypothesis postulates that the civilization created the simulation in order to evolve and develop higher forms of intelligence, safely. The advanced civilization intends to harvest suitable higher intelligences from the simulation once they emerge. These intelligences will escape physical death within the simulated universe and be deployed to perform useful functions in the advanced civilization. Consider an individual agent within the simulation who believes that it would be in its interests to be harvested either as an individual or as part of a larger-scale collective intelligence.

In this example, we will assume that all alternative hypotheses facing the agent are non-beneficial—they will not provide net benefits to the agent if they are acted upon and prove to be correct. In such a case, the agent will maximize its interests if it decides to act on the basis

[1] Bostrom (2003) – see References

of the beneficial hypothesis and spends its life working on itself in order to enhance its intelligence.

Obviously, if the beneficial hypothesis proves to be correct, the agent will be substantially advantaged, continuing to live beyond physical death. Alternatively, it may happen that one of the non-beneficial hypotheses proves to be true. But even if this is the case, the agent will not end up worse off by having acted on the beneficial hypothesis that has proven to be incorrect, provided the costs of doing so are not significant.

Under these circumstances, despite the agent being faced with radical uncertainty, there is a rational strategy for deciding its actions. This strategy will maximise the achievement of its goals.

The agent does not know in advance which hypothesis is true. But nonetheless, it can decide rationally which hypothesis (or class of hypotheses) it should act upon, as if they were true.

There are many variations on this theme. For example, an agent might face a mixture of plausible hypotheses that produce different combinations of benefits and harms. The field of decision theory studies these more complex cases. It sets out to identify what a rational agent should decide in various circumstances, given their particular goals, even in the face of radical uncertainty.

Different agents might pursue dissimilar goals. The decision-making strategies that are optimal may vary for different goals. For example, decision theory evaluates what a rational agent should decide if its goal is to maximize its benefits, or to minimize harm to itself, or to balance risks and benefits in some defined way, or to minimize the maximum regret that it might experience, and so on.[2]

The existence of decision theory and its rigorous insights enable us to go beyond the reach of current science. The application of its discoveries enables us to make rational decisions about how we should live our lives, even though we face radical existential uncertainty.

What specific implications do these methods have for deciding how we should act now, given possible explanations for the existence of the universe? Fortunately, we do not have to deal with an infinitely huge number and diversity of classes of plausible hypotheses. This is because evidence is available that reduces the number of hypotheses that we need to consider. This evidence significantly decreases the number of hypotheses that can be considered to be plausible.

In particular, it is evident that we live in a universe that is consistent with the hypothesis that it is fine-tuned in many ways to be life-friendly. Furthermore, the fine-tuning is not conducive only to the emergence of simple life. It also appears to give rise to an evolutionary process with a particular trajectory. As we have seen, this trajectory results increasingly in the integration of living processes and in enhancing their evolvability/intelligence.

As the trajectory unfolds, cooperative organisations of greater and greater scale emerge, and as the scale of these increases, so too does their evolvability/intelligence. If this trajectory continues to unfold successfully on Earth, the intelligence of the emerging global entity will

[2] e.g., see Bejleri *et al* (2022) – see References for full citation

far surpass the intelligence of any individual human. Nonetheless, individuals will contribute to the intelligence of the global entity, just as our brain cells contribute to our intelligence.

Many possible hypotheses about why our universe exists are not plausibly consistent with its apparent fine-tuning. We can exclude them from further consideration. In general, this includes all 'right-hand path' spiritual and religious traditions that reject the enhancement of agency as a central goal of their practices and beliefs.

Of the remaining hypotheses that retain their plausibility, there are some that would provide benefits to individuals if they act in particular ways during their lives. I have already briefly discussed examples: our universe might have been intentionally set up to produce higher levels of intelligence. It may be that simulating such a universe is the only way to grow higher intelligences safely. It is possible that intelligences that emerge within such a universe will be advantaged if they contribute to the further development of themselves and of other intelligences in their universe.

For example, they may be part of intelligent processes that are subsequently harvested and used for functions outside the simulated universe, or they may perform permanent functions within the simulated universe. Alternatively, an advanced civilization might have set up a fine-tuned simulation in order to provide an optimal environment for the further development of individuals from within the civilization itself.

But the evidence that suggests our universe might have been intentionally fine-tuned also leaves us with many other hypotheses. These include ones that would not provide any existential benefits to individuals who emerge within such a universe, no matter what they do or contribute. For example, the creators of the simulation might have produced it simply to test hypotheses of their own about how living processes evolve. When their experiments reach an end, they may just discard the simulation and anything that emerged within it.

A further example is the 'multiverse' hypothesis. It explains fine-tuning by assuming that those universes in which life emerges will necessarily exhibit characteristics that enable life to emerge and develop. From the perspective of intelligent life that emerges in such a universe, it will appear to be fine-tuned for life.

Nevertheless, even when an agent cannot tell which of these kinds of universes it lives in, it can be rational for the agent to act 'As If' the predictions made by the most beneficial hypothesis are true, all other things being equal.

These broad considerations provide strong and rational reasons for the adoption of life-goals that are pro-evolutionary. This would involve doing what you can to advance the trajectory of evolution by developing your own intelligence/evolvability, as well as contributing to the development of the intelligence/evolvability of the larger-scale cooperatives in which you are embedded. On this planet at this time, a priority is to contribute to the emergence of a cooperative and highly-evolvable global society.

Of course, there are no absolute guarantees that embracing an evolutionary worldview will pay off for individuals in the longer term. But there are no downsides associated with doing so. If fact, there are many benefits: as you work on yourself to enhance your evolvability, you will no longer be harmed by negative emotions and feelings and will be able to experience satisfaction and joy in whatever you choose to do. You will also develop enhanced cognitive capacities that will enable you to generate effective strategies for achieving your goals, whatever they may be.

In summary, if you wish to live your life in a way that maximizes your existential interests, you should spend your time contributing positively to the purposes for which our universe might well have been set up. If this proves to be futile, as it may, it seems that there are no plausible alternative strategies that are likely to produce a better outcome for you.

However, I have only sketched a broad outline of the relevant considerations here. This is not the place to present a far more comprehensive model of the strategies that a rational agent should adopt in the face of radical uncertainty about why there is something rather than nothing.

Furthermore, like science itself, the conclusions reached by this kind of approach are likely to change as assessments of plausibility evolve and as more evidence emerges. New evidence might open new plausible hypotheses. Other kinds of fresh evidence might rule out some of the hypotheses that are now considered to be plausible and consistent with current evidence.

As far as I know, only one thinker has used this kind of approach previously in a serious attempt to address the big existential questions. French mathematician and philosopher Blaise Pascal used a similar line of reasoning to argue that a rational individual should act as if the Christian god actually exists, and follow the tenets of Christianity. If the individual does so, and if there actually is such a god, the individual will go to heaven and avoid hell. If the alternative proves true, the individual will lose little.

However, Pascal's approach has several obvious flaws. Foremost among these is that the only beneficial hypothesis that he considered plausible was that the Christian God existed. He did not include in his analysis any of the other religious traditions that are at least as plausible as his version of Christianity (and are equally implausible in the context of modern knowledge).

Broadly, this kind of thinking led me to decide that I will endeavour to endure and learn from any suffering that comes my way in this universe. As has been the case for many of the misfortunes in my life up until now, this suffering may provide important learning experiences and motivation for my future development.

Furthermore, if I can develop my capacity to accept fully whatever circumstances arise, I will become more like the Buddhist monk on the steps of the United States embassy in Vietnam. If I get there, there will not be any downside: whatever suffering comes my way, I will not flinch or react negatively. The experience will be no different from any other set of bodily sensations. Furthermore, as I have found, once suffering has passed, it is as if it never happened.

Of course, the idea that life in this universe might be a temporary experience that is intended to provide learning experiences is common across the great spiritual and religious traditions, as well as in New Age 'thinking'. These sources variously suggest that we are reincarnated until we develop sufficiently; that we are 'spiritual beings having a human experience'; that we are 'the one's' way of experiencing a greater range of possibilities; and so on. But as with most beliefs within the spiritual and religious traditions, their explanations and justifications are often mutually contradictory. Their beliefs need to be shorn of their spiritual mumbo-jumbo, and understood instead using rational approaches.

These realizations had a major impact on my attitudes to life and its challenges. Previously, I had fallen into the trap of concluding that if science was unable to understand why our reality exists, it was pointless to consider the issues any further.

This led me to conclude that any thinking about possible explanations for existence was senseless and futile and could not possibly go anywhere. Science was our best method for working out how the world works, and if things were beyond the reach of science, they were beyond rational knowing.

But this thinking prevented me from seeing the implications of another obvious possibility. Although we do not have the science-based tools to account for why our reality exists and probably never will have those tools, this does not bring our explorations to an end. It does not prove that there is no possible explanation for our existence that can make sense of our lives. It suggests only that if there is such an explanation, we are incapable of demonstrating its validity during our lives. But all this means is that we cannot rule out any particular possibility. For example, it does not enable us to rule out the possibility that when our bodies die, our intelligence and being may continue to exist in some other realm. Neither can we rule out the alternative.

We simply do not know.

We cannot know what death will bring us. But this means that we will have no reason to be surprised if one of the more 'hopeful' possibilities proves to be true. Consequently, the absence of certainty about these possibilities should not determine how we enter the dying process. It is consistent with all scientific knowledge and other facts to enter the dying process with unrestricted curiosity about what, if anything, comes next. As the voice told me when I took my heroic dose of mushrooms: "Very clever, John. But you don't know shit about what is really going on here." I now take this message to be true, but in a positive sense. What is really going on here might make additional sense of my existence.

Furthermore, as I have outlined, there are ways in which we can live here and now that will pay off in particular kinds of universes—for example, in universes that provide us with continued existence if we work on developing ourselves and the larger living systems in which we are embedded. Broadly, this entails aligning our lives and actions with the trajectory of evolution. The evolutionary worldview identifies in detail what this requires.

These considerations led me to a conclusion that was not very different from the position I had reached before my bout of depression. But the context in which I reached it has changed radically. I am now very much open to the possibility that what is going on here in our universe makes sense in some larger context that includes but transcends this universe.

At the end of all my explorations, I arrived back where I had started, but understood its wider implications for the first time.

As I have indicated, this has fundamentally changed my attitude to suffering and death. I still do not know with any certainty what is really going on here. But now I know that it is possible to make rational decisions about things beyond the reach of science and about which I have no certain knowledge.

Ultimately, it may prove existentially futile to live my life as if the values that seem to be implicit in our universe also make sense in a larger scheme of things. But amongst the many possible hypotheses that could account for why there is something rather than nothing, the pay-off if such a hypothesis turns out to be correct is potentially very significant.

* * *

However, I was unsettled by the conclusion that this new kind of rational thinking could deal effectively with issues that apparently fall outside the reach of science.

Up until this point, my thinking had proceeded on the assumption that scientific investigation, broadly understood, is humanity's best method for discovering how we should act in the world to achieve our goals, whatever they may be. However, I had now developed a new, rational method for deciding how to act in circumstances that could not be understood using scientific methods. This new approach seemed to be a reliable and rational decision-making method.

Was it possible to unite the two forms of thinking in a new synthesis? In such a synthesis, each of these forms of thinking would be seen as instances of a single, higher-level framework. The new framework would embody a unified approach that subsumed both of these forms of thinking. It would identify rational methods for making decisions that worked effectively in both domains—for science and for circumstances characterised by radical uncertainty.

The potential for such a possibility seemed to me to be heightened by the fact that the methodology of science has never been properly established. Science has never demonstrated from first principles how its methods overcome the problem of induction and the associated problem of causation. Nor has philosophy.

The great Scottish philosopher David Hume demonstrated the seriousness of these problems. Broadly, he pointed out that when a particular event A occurs, and when event B has always been found to follow event A, this does not prove that A causes B or even that the next time A occurs, B will occur again. In these circumstances, it does not even prove that B will occur with greater probability.

What Hume argued is incontrovertible. No matter how many times event A is followed by event B, it does not rule out the possibility that the next time A occurs, B will not.

Science survived this devastating attack on its foundations because of its apparent success at discovering laws of nature and harnessing them to satisfy human goals. There is abundant evidence that science has worked brilliantly in practice, despite Hume's arguments.

Science's response to Hume has not been to disprove his arguments. Instead, it has been to just get on with science and continue producing extremely valuable discoveries.

Scientific methods have been claimed to discover many regularities and laws. Generally, these regularities continue to manifest whenever their existence is tested by observations. But of course, this does not mean that the apparent past regularities will hold true in the next minute or hour. Furthermore, as Hume pointed out, the fact that scientific methods appear to have worked to some extent in the past, does not mean that they will continue to work in the future.

Nor is it the case that the more times that an event B occurs immediately after A occurs, the more probable it is that B will be found to follow A in the future. Probabilities cannot be rationally assigned to future events on the basis of past experiences. Hume's argument applies equally to assigning probabilities as it does to predicting certainty.

In any specific case, it is impossible to rule out the possibility that our universe might be arranged in such a way that in the next instant, all previous apparent laws of nature will cease to apply. Or the possibility that any one, or any combination of them, will cease to apply.

The possibility cannot be ruled out that apparent laws may be replaced at any instant by any one of an infinite number of alternative laws that would produce entirely different outcomes. From this instant onwards, this may occur every second, or millisecond. There are endless possibilities. None of them are ruled out by the observation that up until now, the rules of nature that we believe to have discovered in the past, appear to have continued to apply.

Furthermore, there is nothing that enables us to say with any validity that any new laws that emerge will continue to apply in the future. Nor can we demonstrate with any certainty that the longer a new law applies, the more probable it is that it will continue to apply. And so on, and so on.

Karl Popper's ideas about the philosophy and methods of science are perhaps the most widely accepted within science. He agreed with Hume that observations that are consistent with the predictions of a scientific hypothesis can never demonstrate that the hypothesis is correct.

However, Popper argued that observations that are inconsistent with the predictions of a hypothesis can falsify or disconfirm it. He argued that contrary to naïve inductionism, science progresses not through the verification and confirmation of hypotheses, but by the refutation of 'wrong' hypotheses. Logically, hypotheses can be disproven, but never proven. According to Popper, science progresses through the accumulation of hypotheses that have survived all attempts to falsify them, and through the rejection of falsified hypotheses.

But did Popper's approach really overcome Hume's arguments? Did it actually put science on sound foundations?

No, it did not. As Hume might point out, any number of failed attempts to falsify a hypothesis does not prove that in the next instant in time, new attempts to falsify it will also fail. The predictions of hypotheses that were falsified in the past may make predictions that prove correct at any instant in the future. This and an infinite number of other possibilities cannot be ruled out in each new instant, no matter what possibilities actually occurred previously. At every future instant, an infinite number of possible hypotheses come back into possibility, including those that have been falsified previously.

Is it possible to find some other way in which science can be placed on sound foundations? The answer to this fundamental question depends on what we take to be meant by 'sound foundations'. If we are asking whether science can arrive at certain knowledge about the future, the answer is clearly no. Throughout history, science and philosophy have stubbornly sought a scientific methodology that can be proven to be capable of discovering truths that will apply in the future. But this pursuit has failed. Science and philosophy have been relentlessly exploring a blind alley.

There can be no such proof. Hume's arguments demonstrate this. Science and philosophy have been asking the wrong questions.

The way through these difficulties is broadly the same as I developed for dealing with the problem of existence. It is founded on the realization that an agent can act in ways that qualify as rational, even when confronted with radical uncertainty. Here, rational decisions are taken to be those that can be demonstrated to increase the likelihood that the agent will achieve its goals. Radical uncertainty is faced by an agent in relation to a particular decision when forecasting is impossible, and the agent is unable even to assign probabilities to the possible outcomes of its actions.

As we have seen in relation to existential uncertainty, even when facing radical uncertainty, an agent can still arrive at a decision using a method that will maximize the achievement of its interests, provided particular conditions are met.

The example of existential uncertainty that we explored considered circumstances in which a particular subset of possible outcomes existed amongst a much larger set of possibilities. For this subset, if the agent acts in a particular way, and if this possible outcome actually occurs, the agent will benefit. Provided that the downsides experienced by the agent are small when it acts in this way, and there are no other possible outcomes that would benefit the agent, it is in its rational interests to decide to act in this particular way. Such a decision will increase the likelihood that it will achieve its goals.

In these circumstances, this kind of decision-making strategy will always produce the best possible outcomes for the agent, no matter what particular outcome actually arises.

How can this kind of reasoning be applied to the foundations of science? As we shall see, it is applicable because science is faced with radical uncertainty. We will begin our analysis by considering an agent that fully accepts Hume's argument. As such, when the agent makes decisions about how it might act in order to advance its future interests, it knows that it faces radical uncertainty. It accepts that it is faced with an infinite number of possible outcomes in the next instant. It has no rational basis whatsoever to assign different possibilities to any of these possible outcomes. What outputs from science, if any, should the agent use in making its decisions? How should it act to achieve its goals optimally?

In answering these key questions, it is useful to begin by considering the possible outcomes for an agent if it makes its decisions on the basis of hypotheses that conflict with those established by science. In other words, it makes its decisions using the predictions of any of the infinite number of hypotheses that are inconsistent with the laws and other regularities that have been established by science.

The key point is this: Even if the predictions turn out to be correct, the agent may not benefit at all. This is because, if the universe suddenly begins to behave in a way that is inconsistent with the way it has in the past, life in the changed universe may no longer be possible. An agent who is extremely lucky to choose the correct one of an infinite number of possibilities may not survive the experience.

As I have mentioned, there is a lot of evidence that the laws and initial conditions of our universe appear to be fine-tuned in ways that enable life to arise and persist. If the regularities that underpin these laws were suddenly to cease to exist, the universe may cease to be life-friendly. Furthermore, evolution has taken advantage of these laws and regularities to build the processes that enable living processes to function effectively. Consequently, the continued functioning of our nervous, physiological, metabolic, cellular, and other processes depends on the continued existence of these regularities. If these past regularities were suddenly to cease to exist, so too would our lives.

Our science is not sufficiently advanced to identify in detail which of these past laws, regularities, and other patterns are essential to our continued existence. But the complex interrelationships of the constituents of our bodies and those between our bodies and our environment suggest that even slight changes to past patterns may be fatal, instantaneously.

Given these considerations, if a living agent bases its decisions on hypotheses that are inconsistent with past patterns, and if the predictions of these hypotheses suddenly emerge, the agent is unlikely to survive to appreciate any benefits from getting it right.

There is only one decision-making strategy that an agent can adopt that will ensure that it will benefit at least in some cases, no matter what hypotheses prove to be accurate. This strategy is to base its decisions on the meta-hypothesis that regularities that appear to have prevailed in the past will generally continue. Of course, a community of rational agents will hedge its bets somewhat by including research programs that test the boundaries of this hypothesis through time.

These considerations suggest that the rational reason why we should base our decisions broadly on the predictions of established science is not because they are more likely to be correct. Rather, it is because even if hypotheses that conflict with past patterns prove to make accurate predictions about the future, relying on these hypotheses will not get us anywhere. If they are correct, it is likely that we will not be around to enjoy any benefits that might flow from our accurate predictions.

This is not the place to develop in detail all the consequences and nuances of this proposed approach to the philosophy and epistemology of science. Suffice it to say that the arguments I have sketched justify scientific methodologies that are broadly similar to those practiced by the scientific community. There are important differences in detail, but it is beyond the scope of this book to consider these in greater depth.

Largely, science has been doing it right, but for entirely the wrong reasons.

These insights point the way toward the development of a rational approach to making decisions that applies equally to the methodology of science as well as to fundamental existential issues. However, the specific approach I have taken to making decisions that rely on scientific methodologies cannot be applied to decisions about events that occur after the agent has died. The methods that are specific to science cannot be applied in circumstances where the testing of hypotheses is not possible.

However, at a higher level of abstraction, the same kind of reasoning that was used in relation to science is equally applicable to decision-making about consequences that are not experienced until after the agent dies. Both cases deal with how an agent can make decisions rationally, even in the face of radical uncertainty about the future consequences of decisions.

Any conclusions reached by this reasoning in either domain will have equal validity. Conclusions reached about how an agent should act now in the face of post-death possibilities are no more nor less justified than science-based conclusions about how to act in relation to pre-death events. At a higher level of generality, this approach unifies science with existential reasoning. In neither case does it generate certain knowledge about future events. However, in both cases, it identifies how an intelligent and rational being should reason in order to undertake decision-making in the face of radical uncertainty.