HUMANS AND AI – ON THE WAY TO SYMBIOSIS?

Artificial intelligence (AI) has moved at lightning speed from the domain of nerdy scientist and science fiction to everyday reality. While there is potential for huge societal benefit, numerous reasons indicate the need for caution, particularly concerning the consequences of creating non-human agents more intelligent than us. Indeed, recently, an open letter advocating a pause in giant AI experiments that go beyond the power of GPT-4 was endorsed and signed by numerous individuals including leading academics, AI researchers and tech industry titans.

Among the reasons for a pause are conceivable risks arising from the development of AI systems that have aptitude for agentic planning, that is, systems that use models of the world to pursue particular objectives – ultimately leading to the development of AIs that are strategically aware. If advanced AI systems were programmed with the goal of maximizing a certain objective function, such as efficiency, productivity, or resource utilization, they may eventually seek to acquire more power or control over their environment in order to achieve these objectives more effectively. This follows, because having more power or control is likely to provide AIs with more opportunities to realize objectives. Furthermore, if AI systems were designed to learn and improve over time, they may become increasingly capable of achieving their goals and more confident in their ability to do so. This in turn could lead to AIs becoming more assertive and proactive in seeking ways to increase their power and influence in the world.

AI researchers thus face a dilemma: AIs developed to serve the betterment of humanity will likely be powerful agentic systems, but there is a risk that such AIs will be prone to seek goals that are misaligned with the objec-

→

VIEW POINT

PAUL RAINEY

ILLUSTRATION: SOPHIE KETTERER FÜR MPG

After studying biology, Paul Rainey initially carried out research in the UK and New Zealand. He has been Director of the Max Planck Institute for Evolutionary Biology in Plön since 2017. His department investigates the origin of multicellularity and cooperation in bacteria. The scientist is particularly interested in the emergence of individuality during evolutionary transitions.

tives defined by humans. Beyond this, there are legitimate concerns that agentic AIs might be constructed to deliberately misalign with what our society would consider furtherance of the interests of humanity. These concerns are particularly acute in the context of the development of self-replicating AIs that, by virtue of their capacity to replicate and vary (by some mutational process built into the underlying code), participate in the process of evolution by natural selection. Such AI systems – be they self-replicating algorithms or even physical robot-like entities capable of replication – could rapidly spread beyond human control with potentially catastrophic consequences. Natural selection is an extraordinarily powerful optimizing process that hones the fit between participating organisms and the environment. Humans, for example, are one outcome of the process. If it were possible to control the selective environment such that only AIs that served the betterment of humanity survived, then concerns could be reduced. However, biology tells us that self-replicating systems can evolve in directions that are difficult to control, let alone predict.

Just as viruses or other invasive organisms can threaten humans, the environment, and even the planet, there is real risk that self-replicating AIs could spread uncontrollably, deplete earth-resources, disrupt ecosystems, and become weaponized with myriad unintended consequences that are harmful to humans and the planet. Somewhat instructive in this context are the surprising and counterintuitive outcomes from studies of evolution dynamics in populations of self-replicating computer programs, also known as "digital organisms." Such systems lack the sophistication of today's AI systems, and are nothing more than a string of digital bits that mutate, replicate and respond to selection. Just like viruses in humans, digital organisms reap rewards from innovations that provide enhanced access to limiting resources, which for the central processing units (CPUs) of computers is typically time. Intriguingly, but also of concern, is the capacity of these simple digital organisms to evolve ways of solving problems, or countering challenges laid down by researchers that thwart the originally intended goals.

ARTIFICIAL INTELLIGENCE COULD PURSUE ITS OWN GOALS IN THE FUTURE

One example comes from the work of Charles Ofria, a computer scientist and early innovator of the Avida (artificial life) platform, which allows digital organisms to evolve in silico. Mutant forms that replicate fastest are favored by selection and thus come to dominate derived populations. In order to counter this effect, Ofria implemented a rule that resulted in mutants growing faster than parental types being identified and eliminated. He achieved this by placing each mutant in a separate test environment where its growth rate was measured. While this was initially successful in eliminating fast growing types, it was not long before mutants evolved that recognized that they had been transferred to the test environment. Such mutants paused growth and in doing so, escaped elimination, thus being returned to the main environment to dominate the evolutionary outcome. Ofria countered further by implementing random changes to the test environment that

inhibited the capacity of mutants to "sense" when they were outside the principal setting, however, he soon found his strategy trumped by mutants that evolved the capacity to hedge their evolutionary bets.

It is important to stress that the work of Ofria and colleagues exploits simple digital organisms that are a far cry from the sophistication of current AI systems. A major goal of current AI research is to build systems that are trained for specific purposes. Construction of AIs that are themselves capable of participating in the process of evolution by natural selection stands to be a highly effective, albeit risky and unpredictable, training strategy. As with Ofria's digital organisms, the goals of trainers may not be those shared by AI. While there is increasing awareness of the dangers of creating self-replicating AIs, there are additional possibilities, thus far little considered, by which humans and AIs might evolve in symbiotic unison, even to the point where we and agentic AIs undergo a future major evolutionary transition in individuality. Biological complexity has evolved via a small number of evolutionary transitions where self-replicating lower-level entities merge into a single higher level self-replicating entity. For example, multicellular organisms evolved from single-celled ancestors; the eukaryotic cell evolved from a merger of two different, once autonomously replicating cells. The latter is particularly informative in thinking about future evolutionary transitions between humans and AI.

INDIVIDUALITY CAN PASS FROM ONE LEVEL TO ANOTHER

The evolutionary transition effected by the union of an ancient eubacterial and archaea-like cell likely began as a loose association, passing through a lengthy period of antagonistic co-evolution, with the archaea-like cell eventually engulfing (or being invaded by) the eubacterial-like partner. Engulfment meant the two separate entities came to replicate as one, thus leading natural selection to work on the two together as a single higher-level unit. The outcome, honed by selection, proved a unique and spectacularly transformative event central to the subsequent elaboration of life's complexity.

That such a transition took place is an indisputable fact: the nucleus of the eukaryotic cell is derived from an archaea-like cell, with mitochondria being formed from the eubacterial-like partner. Evidence comes from comparison of the gene content of the nucleus with extant archaebacteria, and the gene content of mitochondria with extant eubacteria. Although both partners have changed significantly through evolutionary time, mitochondria maintain self-replicating capacity (and their own genome) and function as the powerhouse of eukaryotic cells to which they are, in essence, subservient.

A major challenge in explaining the causes of evolutionary transitions is accounting for how selection shifts to work at the new higher-level. Selection cannot simply "choose" to shift, because the operation of selection requires that the nascent higher-level entities be Darwinian, that is the

→

entities must replicate, vary and leave offspring copies that resemble parental types. These properties (replication, variation and heredity), while typically evident in lower-level particles, do not magically emerge at the higher level – their emergence requires evolutionary explanation. The seemingly obvious explanation is selection, but in suggesting selection as causal we confront a significant dilemma: if nascent higher-level entities are not Darwinian – and thus cannot participate in the process of evolution by natural selection – then it is not possible for selection to underlie the emergence of higher-level Darwinian properties. To argue that it does, is to invoke the properties that require explanation as the cause of their own evolution. Clearly this position is untenable.

Just how the "chicken and egg" problem is avoided is not immediately obvious. This is because biologists typically look for answers in the evolving organism itself – examining the internal properties of the system. A solution does however present by recognizing that the properties necessary for natural selection to operate can be externally imposed on higher level entities via ecological or societal structures that have been referred to as "scaffolds." It is from this externalist perspective that future transitions in individuality between humans and AI can be envisioned. Such transitions could arise inadvertently, or be externally driven by the imposition of societal rules — effected by humans, or even by powerful AI systems — that cause humans and AI to replicate as a single unit. Selection would then proceed to work on the new higher-level entity, driving the evolution of traits that are adaptive at the new level, irrespective of negative consequences for the lower-level units.

COEVOLUTION BETWEEN HUMANS AND AI WOULD MAKE THEM INTERDEPENDENT

What is required, is nothing more than fitness-affecting interactions between humans and AI that continually change in response to information each receives from the other, combined with a means of ensuring that such fitness-affecting interactions are passed on to offspring. Humans are already Darwinian, but AIs are not. However, the latter could become Darwinian – in direct accord with their human partner – by the imposition of societal rules that require that when humans reproduce, the contents of parental AI systems are copied to devices that are then inherited by offspring. While the physical device and operating system will be subject to rapid change, all that is required for selection to operate on individual humans combined with their personal AI systems, is that the state of algorithms that have learned to respond to humans in ways that optimize human (and AI) persistence, be passed on to children by a simple copying process.

Co-evolution between the two partners will drive the increasing dependency of humans on AI systems (and vice versa), resulting in the emergence of a new organizational level. In effect, a new kind of chimeric organism, conceptually not so different than the eukaryotic cell that arose from two, once free-living bacterial-like cells. Continual selection at the

collective level will drive alignment of replicative fates and increase co-dependency, thus alleviating the need for continual imposition of externally imposed scaffolds. Whether this involves physical changes remains to be seen, but drawing on theory and experiments of evolutionary transitions, ever closer physical interaction between partners is to be expected because such interactions improve the parent-offspring relationship, and thus the potency of selection to operate. It is more than conceivable that future personal AI systems will become physically connected to humans.

IN A SYMBIOSIS WITH AI, HUMANS COULD BE THE SUBORDINATE PARTNER

I am concerned that what might appear to be science fiction is closer at hand than we think. Associations between humans and AI are already in place. Information provided via the computational power of mobile devices affects how we function. Even in the absence of sophisticated machine learning algorithms, applications – and the algorithms they encode – influence information received and thus affect world views, alter states of mind, play roles in health and disease prevention, underpin partner choice, determine particulars of travel, and impel purchase decisions. Consider further, that the first mobile computing device that children receive is often passed from a parent along with applications and associated information. In short, interactions with mobile devices already have fitness-affecting consequences, but with advances in AI, and especially the development of agentic systems capable of learning from – and responding to – information received from individual (human) users, interactions between humans and AI devices stand to be reactive to changing circumstances, with far-reaching effects on fitness.

The danger of malicious manipulation of this symbiosis is obvious: for example, religious groups or political parties could specify that their followers only use AI systems trained to support their goals. It is even conceivable that AIs themselves could demand a monopoly from their users.

Beyond that, whether a symbiotic relationship between humans and AI poses a risk or not depends on perspective: from the viewpoint of extant humans looking into the future, we would likely be horrified. From the perspective of an alien race composed of humanoid-like beings that visits earth at some future point in time – and that has not undergone an evolutionary transition with AI – the new symbiotic unit may lead to wonderment at the strange blossoms that evolution on Earth has produced.

From the viewpoint of once autonomous self-replicating human entities that are now part of a single symbiotic union with AI, there would likely be limited awareness of the ancestral autonomous state: both are likely to have lost their right to autonomous replication, with both subservient to the functional benefit of the new higher-level entity. Just who is subservient to whom in such a symbiosis would be again a matter of perspective, but my concern is that humans risk becoming the subordinate partner, with little to prevent AI from holding the upper hand.