DOI: 10.1111/meta.12595

ORIGINAL ARTICLE

Transdisciplinarity, neuro-techno-philosophy, and the future of philosophy

Nayef Al-Rodhan[1,2,3,4]

[1]St Antony's College, University of Oxford, United Kingdom

[2]Geopolitics and Global Futures Programme, Geneva Centre for Security Policy (GCSP), Switzerland

[3]Institute of Philosophy, School of Advanced Studies, University of London, United Kingdom

[4]Council on Frontier Risks, World Economic Forum (WEF), Switzerland

Correspondence
Nayef Al-Rodhan, Head, Geopolitics and Global Futures, Geneva Centre for Security Policy, Maison de la Paix, Chemin Eugene-Rigot 2D, P.O. Box 1295, 1211 Geneva 1, Switzerland
Email: nayef.al-rodhan@sustainablehistory.com

Abstract

Philosophy and science have always striven to make sense of the world, continuously influencing each other in the process. Their interplay paved the way for neurophilosophy, which harnesses neuroscientific insights to address traditionally philosophical questions. Given the rapid neuroscientific and technological advances in recent years, this paper argues that philosophers who wish to tackle intractable philosophical problems and influence public discourse and policies should engage in *neuro-techno-philosophy*. This novel type of inquiry describes the transdisciplinary endeavor of philosophers, (neuro)scientists, and others to anticipate the societal implications of the impending transformations of subjects and theorizers. While human enhancement is likely to irreversibly change what it means to be human, disruptive technologies might lead to the emergence of artificial intelligent agents and human-machine hybrids. The paper predicts that neuro-techno-philosophy will be indispensable to understanding and engaging with these game-changing innovations and thus play a pivotal role in the future of philosophy.

KEYWORDS

artificial intelligence, future of philosophy, human enhancement, human mind, human nature, machine learning, neurophilosophy, neuroscience, neuro-techno-philosophy, philosophy, public policy, runaway technologies, transdisciplinarity, transhumanism

1 | INTRODUCTION

The future is notoriously unpredictable. The forecasting of academic trends may be guided by an understanding of a given discipline's nature, aim, and past developments. Despite philosophy's long history and philosophers' tendency to reflect on their own discipline, it is surprisingly difficult to establish precisely what philosophy is and what it strives for. As Bertrand Russell argued, any attempt to define philosophy "is controversial and already embodies a philosophic attitude" (1959, 7).

We may improve our grasp of a concept through comparisons and delimitation. In the case of philosophy, it has proven instructive to examine its relationship to other sciences, in particular those involving empirical methods. Philosophy and science are united in their methodological rigor and their desire to understand the world. This methodological and motivational unity gives rise to a dynamic by which progress in one domain acts as a catalyst for progress in the other. Scientific discoveries invite philosophical reflections on their implications for our understanding of the world, just as philosophers' concepts and questions offer insight and inspiration for further scientific research.

Together, philosophy and science shed light on virtually every aspect of our lives. A perennial issue for both disciplines is the investigation of our human nature. Contemporary philosophy and science pursue the maxim "Know thyself" based on a shared acceptance of what I shall call the *naturalized view of human nature*. Two related insights have prompted the shift to this paradigm and, with it, the emergence of new branches of science. The first important insight is the ideological revolution expedited by Charles Darwin's *On the Origin of Species* (1859). This text laid the groundwork for evolutionary biology. Before Darwin, the common understanding—through the millennia from Aristotle to Descartes and Kant—was that humans are fundamentally different from, and hierarchically above, other beings (Mayr 2009). Darwin's evolutionary theory naturalized *Homo sapiens* by placing us as one species in a broader family of all other living things. The second insight underpinning the naturalized view of human nature was produced in the 1950s by the interdisciplinary movement—combining philosophy, psychology, anthropology, linguistics, computer science, and neuroscience—that gave birth to cognitive science (Miller 2003). Neuroscience, which emerged during this period (Cowan, Harter, and Kandel 2000, 344–54), was particularly influential in naturalizing the mind via a physical analysis of mental phenomena. The successful introduction of neurophilosophy in the 1980s demonstrated the value of transdisciplinary approaches. Furthermore, it showcased how the naturalized view of human nature can inform and reform traditional philosophical questions that were heretofore believed to be resolvable only via a priori investigation.

Modern technologies have, however, begun to change the well-established dynamic between philosophy and science. As I argue below, we may expect neuroscientific progress and disruptive technologies to transform radically the core of the project of comprehending ourselves, the world, and our place in it. First, various interventions in the brain will likely enable us to substantially change the human mind and human nature. Second, not only will our own reasoning capacities be improved by human enhancement technologies, artificial intelligent machines have the potential to become reflective entities just like—or better than—us. I argue that these changes to the subject matter and to the competent theorizers demand a novel type of inquiry, which I call *neuro-techno-philosophy*. This innovative conceptual framework will be crucial for philosophers who wish to shape public policy and discourse, which, in virtue of their long-standing tradition of thorough reflection, they are uniquely qualified to do.

I begin my argument with an overview of the neurophilosophical project that unites philosophy and the empirical sciences in the exploration of the human mind and human nature. Next, I outline what I consider to be the most impactful developments in (neuro)science and

disruptive technologies that I expect to change the "what" and "who" of future philosophizing. Finally, I reflect on the relationship between philosophy and science and elaborate on the nature and aim of neuro-techno-philosophy.

2 | THE RISE OF NEUROPHILOSOPHY

Neurophilosophy applies insights from neuroscience to traditionally philosophical questions (Bickle 2009, 4). The term was coined by Patricia Churchland (1986).[1] The field of research inspired by Churchland mainly addresses issues associated with the philosophy of mind, such as the nature of consciousness, the self, and free will.[2] I call this approach the *narrow* view of neurophilosophy. Others, myself among them, use the term "neurophilosophy" to refer to the transdisciplinary research of all disciplines that seek to elucidate the human mind and human nature: although neuroscience is at the forefront, other relevant sciences—including psychology, biology, and the social sciences—are included as well (Al-Rodhan 2021, 1–2, 5). I call this the *broad* view of neurophilosophy.

Some philosophers have been skeptical about the relevance of descriptive, empirical insights for prescriptive, normative issues. This problem was famously articulated by David Hume, who observed a categorical difference between statements about what *is* the case and statements about what *ought* to be the case. Owing to this difference, Hume maintained, the latter cannot be derived from the former without an account of how this gap is supposed to be bridged (2003, 3.1.1.27). There are several ways, however, in which empirical facts bear on normative theories. First, moral, social, and political philosophers often argue for their positions from premises that assume a specific account of human nature. In so doing, they make claims about who we are and what we strive for that may be verified through observation and experiment. Second, the widely recognized is-ought problem is related to the broadly accepted dictum that *ought implies can*. This principle, typically attributed to Immanuel Kant, states that persons cannot be morally obliged to do something that they are unable to do.[3] Neurophilosophy can inform and reform normative philosophy by exploring the neurological, psychological, biological, and other empirically verifiable human limitations. Finally, and more pragmatically, the neurophilosophical project contributes to the development of realistic moral, social, and political theories that take us as we are. For instance, political philosopher John Rawls maintains that knowledge of "the basis of social organization and the laws of human psychology" is indispensable to the procedure of formulating the principles of justice because, without that knowledge, people might fail to be motivated by these principles, in which case "there would be difficulty in securing the stability of social cooperation" (1999, 119).

To illustrate the work of neurophilosophy (broadly understood) and to set the stage for my main argument, let me briefly address four central neurophilosophical questions.

What is the foundation of our moral judgments? The discussion concerning the foundations of our moral judgments is dominated by two opposing views. On the one hand, Hume believed that moral judgments are based on sentiments, which are feelings of approbation

[1]Many credit Churchland with effectively introducing neuroscience to philosophy (Brook and Mandik 2007, 384–85; Bickle 2009, 3). Churchland's position contradicts that of functionalists like Jerry Fodor (1974) and Hilary Putnam (1967), who have insisted on a separation between mind and brain.

[2]Later contributions by Churchland on these issues include Churchland 2002 and 2013.

[3]The *ought implies can* principle has been variously interpreted across philosophical debates. For an analysis of its different uses and a comparison to Kant's original formulation, see Stern 2004.

or disapprobation of character traits. On the other hand, Kant argued that an action's moral worth stems from the fact that the action is motivated by the moral law, which can be recognized though reason. Their respective accounts of moral motivation differ accordingly: Hume held that "reason alone can never be a motive to any action of the will" (2003, 2.3.3.1), whereas Kant claimed that reason can "of itself, independently of anything empirical, determine the will" (2015, 5:42).[4] As Thomas Nagel notes, both philosophers' "influence has been equally great, and . . . contemporary ethical theory continues to be dominated by the disagreement between these two giants" (2012). Today, empirical sciences may provide some indications of which of the two giants was on the right track. Jesse Prinz summarizes research on the relationship between emotions and moral judgments as follows: "Emotions co-occur with moral judgments, influence moral judgments, are sufficient for moral judgments, and are necessary for moral judgments, because moral judgments are constituted by emotional dispositions" (2006, 36). Psychopathology research provides additional evidence for the crucial importance of emotions to moral judgments. For instance, studies suggest that psychopaths show a selectively reduced responsiveness toward sadness and fear, which comes with a deficit in affective empathy (Blair et al. 2001, 2002), and that psychopaths lack a mechanism inhibiting violent reactions in response to distress cues, which is theorized to be causally connected with lacking the ability to distinguish between conventional and moral rules (Blair 1995). These findings indicate that Hume's sentimentalism better describes the nature of our moral judgments than Kant's reason-guided theory.

Is morality innate? This second question concerns the supposed innateness of morality. Are we born with an innate understanding of good and evil, as nativists claim? Or do we acquire a moral sense through learning and experience? Some philosophers have claimed that people are either good or bad by nature, which presupposes innate and intrinsic morality. For instance, Thomas Hobbes (1965) famously argued that humans are naturally ruthless and egocentric, and cooperate with others only by forming a society for their own individual benefit. By contrast, Jean-Jacques Rousseau (2019) held that humans are by nature good and are corrupted only by the influence of society. Whether moral nativists believe human beings to be inherently good or evil, if their assumption were true we might expect our moral judgments and behavior to be robust. Empirical studies reveal, however, that morally irrelevant environmental factors affect our moral conduct and decision-making to a surprising extent. For instance, a study by Thalia Wheatley and Jonathan Haidt (2005) showed that hypnotically induced feelings of disgust influence moral assessments, making individuals incline, ceteris paribus, toward more severe judgments. Using imaging technologies to study the brain functions accompanying moral judgments, Joshua D. Greene and his colleagues (2001) discovered a systematic variation in the extent of emotional processing, suggesting that moral judgments are influenced by the level of one's emotional engagement. Such findings undermine the case for innate morality, indicating the malleability of the human moral compass.

Are we capable of genuine altruism? Neurophilosophy's third central question concerns the debate over altruism and egoism. Although it is indisputable that people sometimes act for the benefit of others, believers in altruism claim that at least certain instances of such behavior are motivated by a genuine concern for another's welfare. By contrast, proponents of egoism argue that all apparently altruistic behavior is ultimately motivated by self-interest. Hobbes seems to support the latter

[4]This dissent should not be taken to imply that Hume dismisses the relevance of reason or that Kant ignores the importance of feelings. For an overview of Hume's and Kant's ethics, see Wilson and Denis 2008.

perspective, stating that "of all voluntary acts, the object is to every man his own good" (1965, 75).[5] In the 1960s, advances in evolutionary biology solidified the dominance of the egoist view.[6] Before that time, group selection theories were invoked to solve a puzzle: why do individuals occasionally act for the benefit of others and at their own expense, at the expense of themselves—seemingly contradicting the logic of Darwin's theory of evolution by natural selection? By popularizing the *kin selection theory*, William D. Hamilton (1963, 1964a, 1964b) offered an egoistic explanation for apparent altruism operating at the level of genes. This theory claims that such behavior may occur when the giver and the recipient of an altruistic act are genetically related, and the advantage conferred to the recipient outweighs the giver's disadvantage, so that altruism raises the probability of the giver's genes being passed on through his kin. With the introduction of the concept of *reciprocal altruism*, Robert Trivers (1971) moreover provided an explanation for apparently altruistic behavior toward nonkin. According to this model, individuals are willing to accept a personal disadvantage to help another if this helping behavior is reciprocated in the long run: the continuous exchange of mutually beneficial actions increases the overall benefit, including the giver's. Provided that theories such as those of Hamilton and Trivers successfully account for all supposedly altruistic acts in terms of self-oriented motives, evolutionary biology seems to vindicate those philosophers who conceive of human nature as egoistic. More recently, the orthodox egoistic view has been called into question. The first of the two most notable challenges was raised in the field of evolutionary biology by Elliott Sober and David Sloan Wilson (1998), who argued that altruistic mechanisms were more reliable and, thus, more likely to have been favored by natural selection than egoistic mechanisms. The second challenge is the one propounded by social psychologist Daniel Batson (2014), whose review of a series of experiments revealed support for his empathy-altruism hypothesis over egoistic alternatives.[7] The work of Sober, Wilson, and Batson reinvigorated the debate on human nature and illustrated the importance of empirical studies for answering neurophilosophical questions.

What are we ultimately striving for? Neuroscience can help us to address this fourth and final question: What ultimately motivates us? The explanation for our drive toward certain goals can be traced back to our neurochemistry. As may be confirmed by even brief introspection, we typically want and like what feels good. Shaped by a long evolutionary process, our brains appear to be preprogrammed to seek out and to enjoy pleasure, including prominently those goods relevant to our survival, like food and sex (Al-Rodhan 2021, 106). Neuroscientific research on addictions has shed light on the workings of the brain's reward system, where drugs stimulate dopamine-producing neurons to generate a strong experience of pleasure. This is the dopamine rush that cocaine addicts, for instance, crave (Nestler 2005, 5–6). Although drug use is obviously a special case of desire-satisfaction, the feeling of gratification triggered by neurochemicals in the brain's reward center also motivates us to repeat other pleasurable behaviors. In other words, we strive toward whatever reliably produces a sense of well-being. This mechanism is also known as the *sustainable neurochemical gratification principle* (Al-Rodhan 2021, 107).

My account of human nature and motivations synthesizes the findings of these four neurophilosophical questions. In my view, the central role that emotions play in moral judgments and cognition, the findings indicating that we do not have an innate morality, and

[5]Bernard Gert has challenged the popular characterization of Hobbes as a "paradigm case of someone who held an egoistic view of human nature" (1967, 503).

[6]Scholars disagree about the popularity of this view. Philosophers like Joel Feinberg claim that the egoistic view was "widely held by ordinary people, and at one time almost universally accepted by political economists, philosophers, and psychologists" (1999, 493; see also Stich, Doris, and Roedder 2012, 147; Batson 2014, 2–3). Others, such as Joshua May, believe it to be "widely and immediately rejected in the philosophical community" (2011, 26).

[7]For an overview and detailed discussion of Sober and Wilson's and Batson's work on altruism, see Stich, Doris, and Roedder 2012, 157–202.

the compelling case for our actions being ultimately motivated by our own perceived self-interest together motivate an account of human nature as fundamentally *emotional, amoral, and egoist* (Al-Rodhan 2021, 63–70). Concerning the question of human motivation, I have offered a substantial account of motives that I take to be particularly powerful and reliable in eliciting neurochemical gratification. We are motivated primarily, I claim, by what I call the *Neuro P5*: power, profit, pleasure, pride, and permanency (Al-Rodhan 2021, 71–81). I return to my neurophilosophical account of human nature and motivations in the next section.

Both the narrow and the broad view of neurophilosophy study philosophical puzzles with empirical insights about the human mind and human nature *as it currently is*. As I have mentioned, however, transformative neuroscience and disruptive technologies are about to drastically change who we are—and new realities demand new approaches. The rise of neurophilosophy as an academic field not only testifies to the productive collaboration between philosophy and neuroscience on deeply meaningful questions but also gives us reason to believe that such transdisciplinarity will serve us well in this upcoming era.

3 | TRANSFORMATIVE NEUROSCIENCE AND DISRUPTIVE TECHNOLOGIES

In addition to providing us with new tools to address old questions, scientific progress and technological development create new challenges. Perhaps neuromodulation leading to transhumanism or runaway technologies superseding humanity still sound like science fiction, but the first steps toward their realization have already been taken. Let me now turn to these issues and outline how I believe these rapid changes are likely to irrevocably transform the subject matter and method of philosophy. Given that the questions explored in the remainder of this paper are prompted by neuroscientific and technological developments, I call the philosophical approach to these changes neuro-techno-philosophy. In the paper's final section, I elaborate on the significance of this term.

3.1 | Human enhancement and transhumanism

Individuals constantly shape and improve themselves in accordance with their aims and values. Familiar examples include the effort we put into our education or into advancing our career, or the intentional formation of habits that we believe will make us healthier and happier. In certain cases, similar improvements can be attained through neuromodulation: the targeted alteration of specific neuronal activities by means of drugs or technological interventions (International Neuromodulation Society 2018). When such interventions aim to augment our abilities, rather than restore them to a given baseline, we refer to them as human enhancement.[8] Means of neuromodulating cognition include, for instance, drugs developed to treat neuropsychiatric disorders and both noninvasive and invasive brain stimulation to improve memory and the ability to focus (Academy of Medical Sciences et al. 2012, 13).

The paradigm shift toward a naturalized view of human nature in modern thought paved the way for accepting that the present features of our nature are not set in stone. Darwin's evolutionary theory undermines any assumption that our species has reached a developmental end point. When the evolutionary perspective of humanity is combined with technological advances and our rapidly improving neuroscientific self-understanding, it suggests that we

[8]For a discussion of different interpretations of the distinction between treatment and enhancement, see Parens 1998.

will develop the ability to manipulate human nature to a substantial degree (Bostrom 2005, 3). Indeed, it is even conceivable that we shall, at some stage, be transformed to the point of no longer being recognizably human but, instead, being *trans*human.

The differences between humans and putative future transhumans raise a number of ethical concerns. Francis Fukuyama argues that our societal achievement of defining fundamental rights is conditional on the existence of an identifiable shared kind—human—to which these rights are ascribed, and to which we all belong.[9] The status of being *more than* human, then, might lead transhumans to claim a more expansive collection of rights (Fukuyama 2004, 42). For this reason, Fukuyama (2002) warns that transhumanism poses a serious threat to our collective sense of a common moral status.

Given that we still do not fully understand the many intricacies of human nature produced by a long evolutionary process, altering our underlying neurological properties may have unpredictable consequences. Fukuyama suggests that, in many senses, the presence of virtues may depend on the presence of vices: "If we weren't violent and aggressive, we wouldn't be able to defend ourselves; if we didn't have feelings of exclusivity, we wouldn't be loyal to those close to us; if we never felt jealousy, we would also never feel love" (2004, 43). If this is right, then neurological interventions aiming to reduce the prevalence of the characteristics we abhor might produce the inadvertent consequence of compromising those characteristics that we value.

Others emphasize cognitive enhancement's potential to bring about radical positive change. Mark Alan Walker (2002) and Phil Torres (2020) explore how such enhancements could fundamentally transform philosophy. Both begin their analyses by observing the apparent gap between philosophy's grandiose aim—discovering the absolute truth about all aspects of life—and philosophers' limited intellectual abilities. Despite our best efforts, certain philosophical problems, like the hard problem of consciousness or the question of free will, remain unresolved—"yet one generation after another strives to reach the false horizons before us" (Torres 2020). Instead of admitting defeat and abandoning all hope of solving such problems, we might, instead, recognize "that it is not we who ought to abandon philosophy, but that philosophy ought to abandon us" (Walker 2002). In other words, the solutions to these philosophical mysteries might lie beyond our abilities, but not the abilities of superior beings. Walker (2002) points out that when faced with the prospect of transhuman philosophers with enhanced philosophical abilities, we are left with two options: we can seize the opportunity to create cognitively superior beings able to advance philosophy in yet unknown ways or we can refrain from doing so. Torres argues that our philosophical interests will be best served by a promotion of mind-expanding technologies: "After all, one way to get a square peg—philosophy's problems—through a round hole—our minds—is to reshape the hole to better fit the peg" (2020).

Regardless of our attitude toward the phenomenon of transhumanism, some maintain that the emergence of transhumans is unstoppable. I count myself among them, as I claim that there is no question *if* it will happen, only *when*, *how*, and *at what cost* (Al-Rodhan 2013). I ground this conviction in my neurophilosophical account of human motivation. As enhancement technologies promise to make it easier to satisfy our desires for power, profit, pleasure, pride, and permanency, I argue that people will almost certainly be highly motivated to develop such technologies and not hesitate to use them. If I am right, then Walker's second option of refraining from using transformative technology will not be viable in the long run, despite ethical worries.

Human enhancement and transhumanism will change (neuro)philosophy in two ways. First, pharmaceutical and technological modifications of our brain's neuronal structures may affect how we feel and think, thereby slowly altering what defines us. Transformations of the human

[9]Fukuyama provides a comprehensive introduction on the potential dangers of transhumanism to society. For further discussions on a variety of philosophical aspects of transhumanism, see Porter 2017.

mind and human nature shift the subject matter of neurophilosophy. Second, insofar as neurological interventions increase our cognitive capacities, technological progress promises to improve our abilities to think rationally, and to recognize and reflect on previously undiscovered connections. Thus, human enhancement has the potential to optimize the theorizer, turning philosophers into transphilosophers.

Not only is technology likely to change humans, it may also play a role in its own right. In the next section, I consider how far machines have already come and what implications their progress brings for our way of doing philosophy.

3.2 | Runaway technologies

In 1950, Alan Turing famously wondered whether machines could think. Recognizing that any purported definition of the terms "machine" and "think" would be contentious, Turing instead asked whether any machine could win what he called the "imitation game." A machine passes this test if, after reviewing sets of answers to written questions provided by a human and a computing machine, an interrogator cannot reliably identify the computer. While no artificial intelligence (AI) to date has been able to pass the Turing test, some believe that a computer will successfully pass it within the next ten to twenty years (Panova 2021). Increasingly sophisticated machine learning has permitted machines to beat humans in other games: IBM's Deep Blue chess computer defeated world champion Garry Kasparov in 1997 (Harding and Barden 1997), and Google's AlphaGo beat the Go world champion Lee Sedol in 2016 (Moyer 2016). While Deep Blue is equipped with an extensive playbook and is preprogrammed to reason in a way that mimics human reasoning, AlphaGo's algorithm depends only on reinforcement learning, meaning it was not provided with any additional information about Go other than the game's rules. AlphaGo became its own teacher, learning from its mistakes and exploring new possibilities with every iteration (Silver et al. 2017, 354). Essentially, the program became a reflective and creative entity. AI's creativity extends beyond mastering sophisticated games, into the realms of art, music, and literature—not only (re)producing humanlike paintings, melodies, and short texts but also creating original art (Miller 2019). Moreover, AI's capacity to recognize patterns and theorize about complex scenarios helps scientists gain otherwise inaccessible insights. Prominent applications include diagnostic medical contexts (Blades 2021) and innovative experimental design, in which AI develops solutions that its creators had not believed possible (Ananthaswamy 2021).

It is a short step, then, from AI's present uses to its future use developing novel philosophical hypotheses and thought experiments. This step, however, requires that machines learn about human values, a task being carried out by the Delphi project. This prototype model for ethical reasoning is trained on a database containing more than 1.7 million moral judgments about everyday scenarios. When asked to assess a given situation, Delphi's answers currently reflect an average American's value system with 92.1 percent accuracy. The researchers' self-declared aim is "to completely close the gap from human-level performance" and "to pave the way towards socially reliable, culturally aware, and ethically informed AI systems" (Jiang et al. 2021, 3). Given the intellectual abilities that AI systems have already demonstrated in scientific contexts, future algorithms might make unprecedented connections in the field of normative ethics, helping us to discover moral principles that human philosophers have, as yet, failed to grasp.

Although it is inherently difficult to predict the future in any context, technological developments are especially unpredictable owing to the dynamic pace of progress. The hypothesized turning point at which disruptive technologies will radically alter our way of life, perhaps including our very nature, is sometimes referred to as the "technological singularity" (Eden et al. 2012, 1). In his "Speculations Concerning the First Ultraintelligent

Machine," Irving J. Good envisions that "a machine that can far surpass all the intellectual activities of any man however clever, ... could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind" (1966, 33).

Given the possibility that machines might become a runaway technology with cognitive abilities that will eventually overtake those of humans, it is important to ask at what point the agency and rights of computing machines ought to be recognized. The standard conception of agency, in the tradition of G. E. M. Anscombe (1957, 2000) and Donald Davidson (1963), is closely linked to intentionality. Whether a machine's performance may be interpreted as an indication of intentionality is, however, notoriously controversial.[10] Instead, I argue that, regardless of the resolution of these debates, we should attribute agency to machines whenever they appear to possess the same qualities that characterize humans (Al-Rodhan 2018, 17). Given my neurophilosophical interpretation of human nature, I suggest that machines ought to be treated as agents if they prove themselves to be emotional, amoral, and egoist.

Prior to the development of machines that can be accurately classified as artificial intelligent agents, it is difficult to anticipate how closely their moral psychology and morality will resemble our own. The question of the similarity of human and machine nature will be relevant to how we deal with these new agents and what we can learn from them. For even if these artificial agents (or strikingly altered transhumans) prove to be intellectually superior to us, they may not be able to fulfill our hopes of answering long-standing philosophical questions about the good, the true, and the beautiful if their conception of what is good, true, and beautiful has nothing in common with ours.

The case for acknowledging artificially created agency is especially compelling whenever the line between human and machine is blurred, for instance through neuromorphic computing. With this brain-inspired technology, neuromorphic engineers replicate the morphology of individual neurons to mimic the neural architecture of the human brain, which could allow machines to share many of the features responsible for the human brain's cognitive abilities (Al-Rodhan 2016). In addition to computer-based attempts to replicate human cognition, scientists are working toward artificially recreating the brain using biological matter. In the laboratory, biomedical researchers have already created so-called human cerebral organoids (HCOs)—miniature, brain-like organs grown from stem cells that have been used as models for understanding cellular mechanisms and for studying pathologies (Lavazza 2021, 2). As scientists continue to create ever more complex HCOs, these organoids might eventually "become a living laboratory for studying the emergence of consciousness and investigating its mechanisms and neural correlates" (Lavazza 2021, 1). At least from the naturalized worldview that believes mental states are in some way dependent on physical matter, it seems possible that an exact—digital or biological—model of the human brain would experience conscious states akin to those of humans.

The prospect of machines gaining agency and cognitively superseding humans has further implications for understanding the world, in terms both of subject matter and of the theorizing entity involved. First, our fellow artificial agents will become an object of interest; their "artificial mind" and "artificial nature" will need to be studied alongside the human mind and human nature. Beyond exploring interesting similarities and differences, it will be important to understand how these different kinds of agents relate to and treat one another. Second, if machines themselves become theorizers, we shall cognitively compete with entities whose intellectual abilities will likely supersede ours over time. Just like

[10]Although John Searle's Chinese Room argument casts doubt on whether mental states can be inferred from an entity's performance (1980, 417–18), he did not show that machines cannot possess intentionality. Rather, he argued that "thinking" cannot be produced by a mere program and, thus, that the Turing test is inadequate. Thinking, according to Searle, requires "machines with internal causal powers equivalent to those of brains" (417).

transphilosophers, artificial philosophers may succeed in solving philosophical issues in, as yet, unimaginable ways.

4 | TRANSDISCIPLINARITY AND NEURO-TECHNO-PHILOSOPHY

I began this paper by outlining how scientific findings have been brought to bear on a range of philosophical questions. Next, I sketched a number of philosophically relevant transformations driven by neuroscientific and technological progress that we may reasonably expect to take place over the coming decades and centuries. In order to address these changes, I call for the development of neuro-techno-philosophy. This approach requires a profound collaboration between philosophers and scientists that differs from currently ongoing collaborations both in nature and in extent. In light of the potential upheavals we face as a society, I predict that neuro-techno-philosophy will become a relevant and consequential part of future philosophy. Moreover, given that philosophers are by training uniquely qualified to reflect on the implications this new era might bring, I also believe that those scholars who wish to shape public discourse *should* engage in this novel conceptual framework to help society understand and navigate what is to be expected. I shall now discuss the understanding of the relationship between philosophy and science that is relevant to the approach, before finally addressing the nature and aim of the neuro-techno-philosophical endeavor.

4.1 | The transdisciplinary pursuit of truth and meaning

When Moritz Schlick, the nominal leader of the Vienna Circle, contemplated the future of philosophy almost a century ago, he anticipated the interaction between philosophy and science that I described at the beginning of this paper. According to Schlick, even though philosophers and scientists have a common aim of understanding the world, they have different approaches to this endeavor. Whereas science sets out to discover the truth, Schlick maintains, philosophy pursues meaning (1938, 126). The scientific method is empirical, meaning that the truth of its hypotheses is tested against actually observable circumstances. By contrast, the philosophical method is purely mental: it is a reflection on actual or possible circumstances, regardless of whether they in fact exist (Schlick 1938, 128–30).

The relationship between philosophy and science has changed over time. In certain historical periods, as in ancient Greece before Socrates, there was no distinction made between the two disciplines. The natural philosophers of these times are the predecessors of philosophers and scientists alike (Curd and Graham 2008). As hypotheses about the world were gradually stated with greater precision, however, and scientific techniques became more sophisticated, sciences such as mathematics, astronomy, and medicine grew to be independent from philosophy (Schlick 1938, 122). Schlick predicted that the same fate would befall other philosophical disciplines, until they all eventually would "become part of the great system of sciences" (132). That since his time the empirical sciences have engaged in greater depth with questions like the nature of consciousness and the foundation of moral judgment can be considered evidence of at least the partial accuracy of Schlick's prediction.

The sciences' increasing interest in questions previously considered purely philosophical does not, however, signal that philosophy is irrelevant or that it is completely decoupled from science. As Schlick emphasizes, the opposite is true. Philosophy, understood as a mental activity that helps us to make sense of ourselves and the world, is an indispensable part of *every* science (Schlick 1938, 130) and of our private lives. By clarifying the meaning of propositions, by asking challenging questions, and by reflecting on new insights, philosophers provide essential

guidance to discourse in the scientific, public, and personal spheres. In this way, the historical divergence of philosophy and science is replaced by a convergence, in which the disciplines are once again united in their quest to understand the world—albeit with a clearer understanding of their respective roles.

4.2 | The nature and aim of neuro-techno-philosophy

Neuro-techno-philosophy describes the transdisciplinary research of all disciplines that seek to understand ourselves and the world, recognizing that this inquiry is about to be fundamentally changed by unprecedented neuroscientific and technological progress. Neuro-techno-philosophers face the prospect that the impending transformation will inform and reform their quest for meaning in two ways. First, it changes the subject matter. Owing to the potential of human enhancement, in a hundred years the study of the human mind and human nature will concern entities substantially different from those it does today. Second, it changes the theorizer. With transphilosophers, artificial agents, and human-machine hybrids on the horizon, future scholars can be expected to use their superior cognitive capabilities to make sense of philosophical issues in a way that is as yet inaccessible to us. Owing to these transformations, the conclusions of past philosophical theorizing may no longer fully apply. Previously established claims, concepts, and theories must be revisited and, in many cases, revised.

Given that neuro-techno-philosophy addresses the implications of future developments, its approach is anticipatory. In contrast to neurophilosophy, which focuses on the human mind and nature *as they are*, neuro-techno-philosophy examines (among other things) the human mind and nature *as they will be*. Owing to the transdisciplinary nature of neuro-techno-philosophy, a thorough understanding of future developments demands a much closer collaboration between philosophers and scientists, with each party having a certain degree of competence in the other's field. Knowledge of our current best science grounds philosophers' pursuit of meaning, just as a familiarity with the method of philosophical reflection enriches scientists' empirical investigation.

Given its intellectual proximity to science and technology, critics might wonder whether neuro-techno-philosophy is still philosophy. Joshua Knobe and Shaun Nichols have pointed out similar controversies concern the status of experimental philosophy (2008, 12). Having emerged at the beginning of the twenty-first century, experimental philosophy applies the techniques of social and cognitive science to study philosophical intuitions. Scholars working in this transdisciplinary research area defend their membership of the philosophical pantheon in various ways. Knobe and Nichols, for instance, claim that the questions about human nature that they analyze are "so obviously philosophical" that it strikes them as self-evident that experimental philosophy is philosophy (2008, 13). If we adopt this criterion, neuro-techno-philosophy similarly qualifies as philosophy: it focuses on classic philosophical concerns regarding human nature, agency, and the conditions of our existence. I suspect, however, that the hesitancy to accept experimental philosophy as genuine philosophy stems not from its subject matter but rather from its use of empirical methods, which places it outside the purely philosophical realm of pursuing meaning. Justin Sytsma and Wesley Buckwalter can be read as responding to this line of thought when they argue that "experimental philosophy *is a way* of doing philosophy" and that its use of empirical inquiry to inform philosophical reflection follows in the footsteps of paradigmatic philosophers such as Aristotle and Hume (2016, 1–2, emphasis added). Although Sytsma and Buckwalter are right to point out that contemporary experimental philosophers are returning to the way philosophy was practiced in the past, this argument does not establish that the empirical parts of the activity of philosophers of the past was indeed itself philosophy. Recall Schlick's claim that when progress had not yet led to the branching out of some of philosophy's subdisciplines,

philosophy and science had formed a unit. Therefore, earlier philosophers' use of empirical methods can equally be understood as doing science *in addition to* philosophy. Similarly, experimental philosophers who complement their reflective activities with systematic empirical studies or statistical analyses can be described as doing both philosophy *and* science. While conducting experiments is integral to experimental philosophy, scientific activity in neuro-techno-philosophy is optional. This differentiation is important to avoid the misunderstanding that neuro-techno-philosophy demands that philosophers change their methods or become scientists. Neuro-techno-philosophy is non-reductive: it rejects the idea that science alone can solve the challenges sparked by transformative neuroscience and disruptive technologies.

The next era, characterized by transhumanism and runaway technologies, calls for urgent foresight in the field of public policy. This task requires highly trained thinkers to help humanity collectively progress in peace, security, knowledge, and prosperity. The importance of this task is why society requires philosophers to embrace neuro-techno-philosophy as part of philosophy's future.

REFERENCES

Academy of Medical Sciences, British Academy, Royal Academy of Engineering, and Royal Society. 2012. *Human Enhancement and the Future of Work*. https://acmedsci.ac.uk/viewFile/publicationDownloads/135228646747.pdf (last accessed: 1 September 2022).

Al-Rodhan, Nayef. 2013. "Inevitable Transhumanism? How Emerging Strategic Technologies Will Affect the Future of Humanity." *CSS Blog Network*, Center for Security Studies, 29 October 2013, https://isnblog.ethz.ch/security/inevitable-transhumanism-how-emerging-strategic-technologies-will-affect-the-future-of-humanity (last accessed: 1 September 2022).

Al-Rodhan, Nayef. 2016. "Neuromorphic Computers: What Will They Change?" *Global Policy Journal*, Durham University, 18 February 2016, https://www.globalpolicyjournal.com/blog/18/02/2016/neuromorphic-computers-what-will-they-change (last accessed: 1 September 2022).

Al-Rodhan, Nayef. 2018. "Artificial Intelligent Agents: Prerequisites for Rights and Dignity." *Age of Robots* 2, no. 2: 11–17.

Al-Rodhan, Nayef. 2021. *Emotional Amoral Egoism: A Neurophilosophy of Human Nature and Motivations*. Cambridge: Lutterworth Press.

Ananthaswamy, Anil. 2021. "AI Designs Quantum Physics Experiments Beyond What Any Human Has Conceived." *Scientific American, Springer Nature*, 2 July 2021, https://www.scientificamerican.com/article/ai-designs-quantum-physics-experiments-beyond-what-any-human-has-conceived/ (last accessed: 1 September 2022).

Anscombe, G. E. M. 2000 (1957). *Intention*. Second Edition. Cambridge, Mass.: Harvard University Press.

Batson, C. Daniel. 2014 (1991). *The Altruism Question: Toward a Social-Psychological Answer*. New York: Psychology Press.

Bickle, John. 2009. "Introduction." In *The Oxford Handbook of Philosophy and Neuroscience*, edited by John Bickle, 3–10. Oxford: Oxford University Press.

Blades, Robin. 2021. "AI Generates Hypotheses Human Scientists Have Not Thought Of." *Scientific American, Springer Nature*, 28 October 2021, https://www.scientificamerican.com/article/ai-generates-hypotheses-human-scientists-have-not-thought-of/ (last accessed: 1 September 2022).

Blair, Robert J. R. 1995. "A Cognitive Developmental Approach to Morality: Investigating the Psychopath." *Cognition* 57, no. 1: 1–29.

Blair, Robert J. R., E. Colledge, L. Murray, and Derek G. V. Mitchell. 2001. "A Selective Impairment in the Processing of Sad and Fearful Expressions in Children with Psychopathic Tendencies." *Journal of Abnormal Child Psychology* 29, no. 6: 491–98.

Blair, Robert J. R., Derek G. V. Mitchell, Rebecca A. Richell, Steve Kelly, Alan Leonard, Chris Newman, and Sophie K. Scott. 2002. "Turning a Deaf Ear to Fear: Impaired Recognition of Vocal Affect in Psychopathic Individuals." *Journal of Abnormal Psychology* 111, no. 4: 682–86.

Bostrom, Nick. 2005. "A History of Transhumanist Thought." *Journal of Evolution and Technology* 14, no. 1: 1–25.

Brook, Andrew, and Pete Mandik. 2007. "The Philosophy and Neuroscience Movement." *Analyse und Kritik* 29, no. 1: 3–23.

Churchland, Patricia S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, Mass.: MIT Press.

Churchland, Patricia S. 2002. *Brain-Wise: Studies in Neurophilosophy*. Cambridge, Mass.: MIT Press.
Churchland, Patricia S. 2013. *Touching a Nerve: The Self as Brain*. New York: W. W. Norton.
Cowan, W. Maxwell, Donald H. Harter, and Eric R. Kandel. 2000. "The Emergence of Modern Neuroscience: Some Implications for Neurology and Psychiatry." *Annual Review of Neuroscience* 23: 343–91.
Curd, Patricia, and Daniel W. Graham, eds. 2008. *The Oxford Handbook of Presocratic Philosophy*. Oxford: Oxford University Press.
Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection: Or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.
Davidson, Donald. 1963. "Actions, Reasons, and Causes." *Journal of Philosophy* 60, no. 23: 685.
Eden, Amnon H., James H. Moor, Johnny H. Søraker, and Eric Steinhart. 2012. *Singularity Hypotheses*. Berlin: Springer.
Feinberg, Joel. 1999. "Psychological Egoism." In *Reason and Responsibility: Readings in Some Basic Problems of Philosophy*, edited by Joel Feinberg and Russ Shafer-Landau, Tenth Edition, 493–505. Belmont, Calif.: Wadsworth.
Fodor, Jerry A. 1974. "Special Sciences (or: The Disunity of Science as a Working Hypothesis)." *Synthese* 28, no. 2: 97–115.
Fukuyama, Francis. 2002. *Our Posthuman Future: Biotechnology as a Threat to Human Nature*. New York: Farrar, Straus and Giroux.
Fukuyama, Francis. 2004. "Transhumanism." *Foreign Policy*, no. 144: 42.
Gert, Bernard. 1967. "Hobbes and Psychological Egoism." *Journal of the History of Ideas* 28, no. 4: 503.
Good, Irving J. 1966. "Speculations Concerning the First Ultraintelligent Machine." *Advances in Computers* 6: 31–88.
Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293, no. 5537: 2105–8.
Hamilton, William D. 1963. "The Evolution of Altruistic Behavior." *American Naturalist* 97, no. 896: 354–56.
Hamilton, William D. 1964a. "The Genetical Evolution of Social Behaviour, I." *Journal of Theoretical Biology* 7, no. 1: 1–16.
Hamilton, William D. 1964b. "The Genetical Evolution of Social Behaviour, II." *Journal of Theoretical Biology* 7, no. 1: 17–52.
Harding, Luke, and Leonard Barden. 1997. "Deep Blue Win a Giant Step for Computerkind." Guardian, Guardian News and Media, 12 May 1997, https://www.theguardian.com/theguardian/2011/may/12/deep-blue-beats-kasparov-1997 (last accessed: 1 September 2022).
Hobbes, Thomas. 1965 (1651). *Leviathan*. Oxford: Clarendon Press.
Hume, David. 2003 (1739–40). *A Treatise of Human Nature*. Mineola, N.Y.: Dover.
International Neuromodulation Society. 2018. "About Neuromodulation." https://www.neuromodulation.com/about-neuromodulation (last accessed: 1 September 2022).
Jiang, Liwei, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. "Delphi: Towards Machine Ethics and Norms." *arXiv*, Cornell University, 14 October 2021, https://arxiv.org/pdf/2110.07574v1.pdf (last accessed: 1 September 2022).
Kant, Immanuel. 2015 (1788). *Critique of Practical Reason*. Revised Edition. Cambridge: Cambridge University Press.
Knobe, Joshua, and Shaun Nichols. 2008. "An Experimental Philosophy Manifesto." In *Experimental Philosophy*, edited by Joshua Knobe and Shaun Nichols, 3–14. New York: Oxford University Press.
Lavazza, Andrea. 2021. "'Consciousnessoids': Clues and Insights from Human Cerebral Organoids for the Study of Consciousness." *Neuroscience of Consciousness* 7, no. 2: 1–11.
May, Joshua. 2011. "Egoism, Empathy, and Self-Other Merging." *Southern Journal of Philosophy* 49: 25–39.
Mayr, Ernst. 2009. "Darwin's Influence on Modern Thought." *Scientific American,* Springer Nature, 24 November 2009, https://www.scientificamerican.com/article/darwins-influence-on-modern-thought1/ (last accessed: 1 September 2022).
Miller, Arthur I. 2019. "Creativity and AI: The Next Step." *Scientific American*, Springer Nature, 1 October 2019, https://blogs.scientificamerican.com/observations/creativity-and-ai-the-next-step/ (last accessed: 1 September 2022).
Miller, George A. 2003. "The Cognitive Revolution: A Historical Perspective." *Trends in Cognitive Sciences* 7, no. 3: 141–44.
Moyer, Christopher. 2016. "How Google's AlphaGo Beat Lee Sedol, a Go World Champion." *Atlantic*, Atlantic Monthly Group, 28 March 2016, https://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/ (last accessed: 1 September 2022).
Nagel, Thomas. 2012. "The Taste for Being Moral." *New York Review, NYREV*, 6 December 2012, https://www.nybooks.com/articles/2012/12/06/taste-being-moral/ (last accessed: 1 September 2022).
Nestler, Eric J. 2005. "The Neurobiology of Cocaine Addiction." *Science and Practice Perspectives* 3, no. 1: 4–10.

Panova, Evgeniya. 2021. "Which AI Has Come Closest to Passing the Turing Test?" Dataconomy, Dataconomy Media, 9 March 2021, https://dataconomy.com/2021/03/which-ai-closest-passing-turing-test/ (last accessed: 1 September 2022).
Parens, Erik. 1998. "Is Better Always Good? The Enhancement Project." In *Enhancing Human Traits: Ethical and Social Implications*, edited by Erik Parens, 1–28. Washington, D.C.: Georgetown University Press.
Porter, Allen. 2017. "Bioethics and Transhumanism." *Journal of Medicine and Philosophy* 42, no. 3: 237–60.
Prinz, Jesse. 2006. "The Emotional Basis of Moral Judgments." *Philosophical Explorations* 9, no. 1: 29–43.
Putnam, Hilary. 1967. "Psychological Predicates." In *Art, Mind, and Religion*, edited by William H. Capitan and Daniel D. Merrill, 37–48. Pittsburgh: University of Pittsburgh Press.
Rawls, John. 1999 (1971). *A Theory of Justice*. Revised Edition. Cambridge, Mass.: Harvard University Press.
Rousseau, Jean-Jacques. 2019 (1750). *The Discourses and Other Early Political Writings*. Second Edition. Cambridge: Cambridge University Press.
Russell, Bertrand. 1959. *Wisdom of the West: A Historical Survey of Western Philosophy in Its Social and Political Setting*. New York: Crescent Books.
Schlick, Moritz. 1938. "The Future of Philosophy." In *Gesammelte Aufsätze, 1926–1936*, edited by Moritz Schlick, 117–34. Vienna: Gerold.
Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3: 417–24.
Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. "Mastering the Game of Go Without Human Knowledge." *Nature* 550, no. 7676: 354–59.
Sober, Elliott, and David S. Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, Mass.: Harvard University Press.
Stern, Robert. 2004. "Does 'Ought' Imply 'Can'? And Did Kant Think It Does?" *Utilitas* 16, no. 1: 42–61.
Stich, Stephen, John M. Doris, and Erica Roedder. 2012. "Altruism." In *The Moral Psychology Handbook*, edited by John M. Doris and the Moral Psychology Research Group, 147–205. Oxford: Oxford University Press.
Sytsma, Justin, and Wesley Buckwalter. 2016. "Introduction." In *A Companion to Experimental Philosophy*, edited by Justin Sytsma and Wesley Buckwalter, 1–2. Chichester: John Wiley and Sons.
Torres, Phil. 2020. "The Future of Philosophy Is Cyborg." *Philosophy Now*, no. 141: 36.
Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46, no. 1: 35–57.
Turing, Alan M. 1950. "I.—Computing Machinery and Intelligence." *Mind* 59, no. 236: 433–60.
Walker, Mark A. 2002. "Prolegomena to Any Future Philosophy." *Journal of Evolution and Technology* 10, Institute for Ethics and Emerging Technologies, Trinity College, https://www.jetpress.org/volume10/prolegomena.html (last accessed: 1 September 2022).
Wheatley, Thalia, and Jonathan Haidt. 2005. "Hypnotic Disgust Makes Moral Judgments More Severe." *Psychological Science* 16, no. 10: 780–84.
Wilson, Eric E., and Lara Denis. 2008. "Kant and Hume on Morality." In *Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), edited by Edward N. Zalta, Stanford: Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/fall2021/entries/kant-hume-morality/ (last accessed: 1 September 2022).

How to cite this article: Al-Rodhan, Nayef. 2022. "Transdisciplinarity, neuro-techno-philosophy, and the future of philosophy." *Metaphilosophy* 00: 1–14. https://doi.org/10.1111/meta.12595