Blog of the APA

A Neurophilosophy of Fake News, Disinformation and Digital Citizenship

By Nayef Al-Rodhan

August 25, 2020

This is the tenth and final post in a short-term series by Prof. Nayef Al-Rodhan titled "Neurophilosophy of Governance, Power and Transformative Innovations." This series provides neurophilosophical perspectives and multi-disciplinary analyses on topics related to power and political institutions, as well as on a series of contemporary transformative technologies and their disruptive nature. The goal is to inspire innovative intellectual reflections and to advance novel policy considerations.

While early fears of misuse of the Internet centered around small-time theft and the ensnaring of minors into inappropriate or dangerous activities, recent years—particularly since 2016—have demonstrated the profound political capacities of the Internet to manipulate popular opinion and influence critical outcomes in the world. The power of "fake news," so-called deep fakes, and systematic disinformation campaigns is now being studied more actively, and, for once, insights from neuroscience have been a significant part of that debate and analysis.

However, in addition to remaining insufficiently understood, fake news is also far from being successfully combated or held at bay. Indeed, the greatest thing that disinformation campaigns have accomplished was to convince many people that there were no disinformation campaigns, thus limiting popular understanding and a willingness to address the distortions that such campaigns result in. Grappling with the true reach and effect of these misleading campaigns and developing a deeper understanding of human vulnerabilities to them constitutes a critical step to limiting their effect.

An important contribution to this understanding in the last several decades has been achieved in psychology and neuropsychology, where the nuances of biases have been more thoroughly appreciated, developed, and articulated. Confirmation bias, among other lesser known biases, prepares our minds to interpret unfiltered information or raw data in alignment with our pre-existing

beliefs and commitments. These findings are powerful and instructive, and yet they too leave under-described the mechanisms by which our neuropsychology operates.

Put differently, these biases studied in psychology are only proximate causes or inputs to human behavior, but our neurochemical make-up and instinctual inheritance underlies these neuropsychological phenomena. It is thus important to 'begin at the beginning' in seeking to better understand human vulnerabilities to the manipulation of knowledge, and two established best practices for interacting digitally in the world.

Fake News

In its contemporary form, the phenomenon of fake news has been profoundly impacted by digital communication channels and social media. The spread of fake news is now faster, more difficult to contain and, ironically, radically 'democratized' – in the sense that it is anyone, anywhere, who can become a creator and messenger of fake news. These elements of novelty *per se* do not make fake news a new phenomenon. In reality, fake news, propaganda and disinformation have accompanied human history and the history of nations and empires. During ancient Rome, at the time of the so-called Second Triumvirate (the alliance between Octavian – the heir of Julius Cesar – and Marcus Antonius) and as the alliance was crumbling and two antagonistic sides were emerging, a ferocious propaganda war followed. Octavian proved to be particularly able in this war, but it was one particular piece of fake news that impacted the course of events. Octavian got hold of a document that he claimed to be Marcus Antonius' will (the authenticity of which is still disputed by historians), which he read in the Senate, turning people in Rome against Antonius. The document allegedly stated that Antonius intended to leave a significant part of his legacy to his children with Cleopatra, the queen of Egypt. Much like fake news today, the document Octavian read was playing on existing prejudices, in this case the anti-eastern sentiments felt by many ancient Romans.

One simple and underlying mechanism of fake news and disinformation has been to reinforce biases and enhance forms of antagonism between groups (often playing on in-group/out-group dynamics). On the one hand, "fake news" is literally the promotion of falsehood, typically through some journalistic outlet. Various examples of propaganda both historic and contemporary fit this definition, as does the new and troubling phenomenon of "deep fakes", wherein sophisticated computer re-imaging introduces very convincing video footage of well-known personalities making statements or carrying out actions that they never said or carried out. This first form of fake news is certainly troubling, but it is importantly distinct from another notion of fake news. In this case "fake news" is an apparently superficial accusation leveled by those in positions of political power at commentators, journalists, and others who issue proclamations, write stories or offer commentary that persons in power object to. This "accusatory" form of the fake news phenomenon is dramatically changing our social and political world.

The greatest trouble with accusatory "fake news," as its perpetrators have long understood, is that the mere establishment of the category in the minds of consumers of digital media threatens the status of truth altogether. It does this by raising the specter of misinformation or disinformation with every item of news reported, naturally also playing upon the psychological tendencies of individuals to disbelieve information which fails to harmonize with their worldview or their desired outcomes in a given scenario (about which more below).

As research surrounding the #MeToo movement has demonstrated, individuals' epistemological confidence can be shaken by the mere suggestion of the possibility of untruth, so that, for example, when an audience listening to a woman confront her accuser is asked "could you be mistaken," or

jurors are told that she might be lying, just this suggestion of a possibility of untruth lessens the level of certitude with which listeners or readers hold their beliefs.

The concept of fake news then, by its very existence, and crucially by its frequent articulation, undermines the strength of belief that citizens have in news up, including science reported in the media. This is one reason why the current cultural moment has been described as a "post-truth era" given that the logical end of undermining epistemological confidence is a situation in which no one is taken to have a viable claim to truth.

The invocation of fake news terminology should thus be considered a threat to transparent and accountable governance, given that the deliberate obfuscation of truth must be understood as the intentional removal of accountability from power. Though in the case of the United States—arguably—the "guardrails of democracy" as referred to by Ziblatt and Levitsky, have held. It is noteworthy that the notion that liberal democracies require "guardrails"—to the extent that they have these and that they are successful—suggests at least some level of understanding of the susceptibility of citizens within these Democratic states to systematic campaigns of disinformation or epistemological manipulation.

Disinformation

From a historical perspective, the tradition of spreading political disinformation is ancient. It is nonetheless clear that the speed of dissemination and effectiveness of focused campaigns of disinformation have achieved what prior political thinkers, theorists, and philosophers could not have imagined possible even a few decades ago. As distinct from fake news, which as we have suggested appears to be an accusation of inaccuracy but in fact is a very deliberate sort of strategy for undermining the status of truth more generally, disinformation is much more specialized in terms of its focus on sowing discord, inciting anger and in some cases violence, and undermining social harmony.

The danger of such campaigns in terms of their success conditions is that they rely upon failures of human psychology when individuals perceive threat, and therefore the campaign need not pay almost any attention to content. Indeed, it is somewhat misleading to describe these activities as "influence campaigns" in any straightforward sense, though they are, of course, ultimately aimed at specific political outcomes. Their success condition was merely that disagreement was exacerbated, political polarization was amplified, and discord was sown within Democratic societies, weakening their overall functioning and making them more susceptible to outside influence.

Neurophilosophy of Bias and Fake News: Primordial Instincts, Neurochemistry and Emotive Responses

The foregoing considerations harmonize with an observation from Yale historian Timothy Snyder, arguing that a propaganda machine can be among the most effective forms of psychological warfare in history. These facts remain sorely under-appreciated, as much of international relations, just war theory, and other fields associated with international conflict still tend to think in terms of visible movements of troops and military machinery or other highly visible displays of aggression and offense. The effectiveness of deploying "fake news" rhetoric and highly selective targeted disinformation, however, is far from accidental. In fact, it is an exhibition of highly refined techniques of sowing social and political discord that to the trained eye should refine and reshape our understanding of what constitutes aggression in the international system.

Why are these methods so effective? Our minimalist evolutionary endowment offers some clues. As I have argued elsewhere, insights from neuroscience demonstrate the extraordinary salience of emotions to human existence, as emotional processing in the brain, on the one hand, and rationality, on the other, are *not* part of dual systems but much more intimately connected. The human amygdala, among others, has been thoroughly studied as part of this mechanism.

Philosophy has long compartmentalized emotions as different from – or rather, as hindrance to – rationality but evidence from neuroscience does not support this claim, just as it does not support the premise of inborn morality/immorality. Coalescing findings from neuroscience, I previously described human nature as *emotional, amoral,* and *egoistic.* Human nature is thus 1. emotional – far more than rational (in fact, emotionality overlaps with a large part of our cognitive functions, including decision-making), 2. amoral – lacking any innate notions of good and bad; this means that human morality will develop in the course of existence and depending on circumstances, although we do have some predispositions, the most powerful of which is for survival, and finally, 3. egoistic – which I conceptualized in the context of this powerful predisposition for survival, which is a basic form of egoism. Therefore, while largely unfinished upon birth, we do not come into being with an entirely blank slate, but rather with a *predisposed tabula rasa*, refined over millennia and geared toward survival (and those acts that maximize our chances of survival). Untutored human nature is therefore complemented only by a narrow suite of survival instincts that promote our welfare. This set of instincts includes pro-social tendencies in concert with the incredible capacity of *homo sapiens* to coordinate their activities for the mutual benefit of all. This inclination for social coordination, however, is sufficiently limited to generate only small groups under most circumstances, with the consequence that "ingroup-outgroup" formation—a readily observable phenomena across a host of biological species—is replicated in human beings at the tribal, ethnic, and national levels. (See also Joshua Greene's book.)

In auspicious circumstances, with positive socialization buttressed by a framework of well-directed education, these tendencies toward narrow grouping can be muted, and the solving of ever larger problems, including those addressed by international organizations, becomes potentially feasible. When, however, divisions, tribal affinities, and polarization are deliberately exacerbated the fragility of these larger cooperative constructions is revealed, and the instinctual response to silo oneself among other "believers" or those most clearly sharing one's group identity can become overwhelming. Disinformation campaigns exploit precisely these vulnerabilities.

From a neurophilosophical perspective, the powerful effect of fake news and disinformation can be framed in the context of older brain structures and the salience of emotionality. The premise that emotions and confirmation bias play central roles in the propagation of fake news comes as an intuitive hypothesis as it is known we will naturally gravitate towards information that is reassuring or that satisfies social and entrenched beliefs. This premise is confirmed by some of the earliest incursions in neuroscience: the prefrontal cortex, which is 'tasked' with our higher decision-making is not always 'switched on' when reading news, but it is rather the social portions of the brain that takes over – and experiments with functional magnetic resonance imaging (fMRI) showed that when reading news that activates the social parts of the brain, the likelihood of sharing that information is exponentially higher. In behavioral terms, the propensity to share information that somehow connects us to a group can be explained simply by the fact that we are more likely to do something if we believe others are also engaged in that behavior.

A related implication is that the same brain regions which validate socially accepted messages will be more resistant to new information that threatens to isolate the individual from the group. A comprehensive study published in *Nature* in late 2019 explored the neuroscience of confirmation

bias, which is the idea that humans discount information which "undermines past choices and judgments". The study provided some unique insights into the fundamental properties of belief formation, aiming to explain both the mechanisms underlying the confirmation bias, as well as how sensitivity to new information is processed in the brain and what it is contingent on.

The question then is if the strength of one's opinions matters in appreciating new information or evidence, generally, or, for instance, if information that disconfirms previous biases is processed slowly and with more scrutiny, due perhaps to surprise. As a practical example, one of the questions the study asked was whether, in particular situations in a courtroom, when a judge considered a defendant to be innocent, it mattered if the prosecutor presented a confident witness that could state otherwise (i.e. that the defendant is in reality guilty), or a less confident one. The study hypothesized that a different level of sensitivity to the strength of others' opinions would most likely be observed in markers of neural activity in the posterior medial prefrontal cortex (pMFC), an area that includes the dorsal anterior cingulate cortex and the pre-supplementary motor area.

Previously, pMFC has been shown to detect "post-decision information" and other studies showed that people with impairments in this brain region may often display "cognitive inflexibility". The conclusions of the study harmonized with 'Orwellian premises' rather than more hopeful expectations of cognitive flexibility. The participants in the study proved less likely "to utilize the strength of others' opinions to re-assess their judgment when it is contradictory", and therefore they were unlikely to alter their judgments when faced with disagreements. In neuroscientific terms, the confirmation bias was demonstrated by "a failure to track the strength of contradictory opinions in the pMFC". Participants, however, were more likely to incorporate the strength of others' opinions when that opinion already aligned with their own.

Behavioral explanations rooted in our social nature – and our craving for social belonging – are, however, not exhaustive. Another perspective, explored both in psychology and neuroscience, has looked at a rudimentary mechanism known as "fluency", which refers to the cognitive ease with which we process information. Information that is repeated on and on appears increasingly truer and in the process easier to comprehend, and thus more comfortable to the brain. A recent study from 2020 explored some of the related processes that help entrench notions of 'truth', with memory playing an important role (as we store all information and past experiences in our memory). However, it appears that memories are used as cues, and we can accept some claims to be true even if they only partially fit with what we remember. In other words, as one ethicist put it, we tend to go with "good enough". In a similar vein, a 2018 study analyzing susceptibility to partisan fake news in the United States concluded that 'laziness' and 'lazy thinking' largely explained the appeal of fake news, even more so than 'bias'. The extensive study revealed that more analytic individuals were better able to differentiate between real and fake news, irrespective of their ideology.

This range of complementary explanations point, crucially, to the many emotional and personal mechanisms involved in assessing new information and forming beliefs. The idea of cold, unbiased, 'mathematical' rationality, while long celebrated in many schools of philosophy, is simply not confirmed by neuroscience, and that is important in understanding both the appeal of fake news and possible ways to fight it.

The Way Forward and Digital Citizenship

Responsible digital citizenship, which implies the responsible use of Internet and digital technology, essentially involves two sets of considerations. The first is geared toward individual safety and well-being when using the mechanisms of the Internet. These considerations are crucial given the internet's

capacity to expose one to various types of surveillance, partially unwitting information sharing, and scrutiny which is otherwise relatively limited in the course of one's normal life. The second set of considerations is that digital citizenship enjoins personal responsibility when consistently and effectively utilizing the tremendously powerful tools made available through the Internet. At a minimum, such citizenship entails not circulating unvetted information and thereby potentially amplifying harmful acts and messages. Beyond this, however, given our now greater understanding of issues associated with fake news and disinformation, and our instinctive biological susceptibility to these sorts of attacks, it is time to acknowledge that digital citizenship cannot be maintained in a state of radical freedom and lack of oversight, and lack of critical cultural awareness.

Instruments to fight fake news have proliferated in recent years, including a wide range of measures such as technical and algorithmic tools for better detection, fact-checking and removal of false content, political and government actions (including in the form of new and dedicated units in ministries and law enforcement agencies), and multilateral initiatives. For example, in 2018, at the European Commission's initiative on "Tackling online disinformation", a Code of Practice on Disinformation was created as a voluntary mechanism of self-regulatory standards to fight disinformation. The code was signed, among others, by Facebook, Google and Twitter. The Commission has carried out monitoring of the implementation of the commitments in the Code. Top-down approaches such as these are critical in limiting the spread of fake news and misinformation, which are of course not only dangerous in elections but in every aspect of life. In the context of the Covid-19 pandemic, *the American Journal of Tropical Medicine and Hygiene* reported that between December 2019 and April 2020, about 5,800 people were admitted to the hospital as a result of false information (rumors, conspiracy theories) spread on social media. Yet, a neurophilosophical understanding of human nature shows that to counter the appeal of fake news requires more than information campaigns – it also requires an appreciation of our emotional amoral and egoistic nature, and our profound need for human dignity. As elaborated in previous posts, human dignity is fundamental to human nature and even more so than the need for freedom. What I mean by dignity is more than the mere absence of humiliation. It includes a set of nine dignity needs: *reason, security, human rights, accountability, transparency, justice, opportunity, innovation,* and *inclusiveness.* In addition, it also requires a greater and more dedicated focus on transcultural understanding and on human civilizations' shared history and interconnectedness. In the absence of transcultural understanding, and given our globalized world, the perils of fake news are poised to become existential.

Nayef Al-Rodhan

Prof. Nayef Al-Rodhan (@SustainHistory) is a Neuroscientist, Philosopher and Geostrategist. He is an Honorary Fellow at St Antony's College, University of Oxford, and Senior Fellow and Head of the Geopolitics and Global Futures Programme at the Geneva Centre for Security Policy, Geneva, Switzerland. Through many innovative books and articles, he has made significant conceptual contributions to the application of the field of neurophilosophy to human nature, history, contemporary geopolitics, international relations, cultural studies, future studies, and war and peace.