Athens Journal of Philosophy

Quarterly Academic Periodical, Volume 3, Issue 3, September 2024
URL: https://www.athensjournals.gr/ajphil
Email: journals@atiner.gr
e-ISSN: 2732-9674DOI: 10.30958/ajphil

Front Pages

Athens Journal of Philosophy

Published by the Athens Institute for Education and Research (ATINER)

****

ATINER is an Athens-based World Association of Academics and Researchers based in Athens. ATINER is an independent and non-profit Association with a Mission to become a forum where Academics and Researchers from all over the world can meet in Athens, exchange ideas on their research and discuss future developments in their disciplines, as well as engage with professionals from other fields. Athens was chosen because of its long history of academic gatherings, which go back thousands of years to Plato's Academy and Aristotle's Lyceum. Both these historic places are within walking distance from ATINER's downtown offices. Since antiquity, Athens was an open city. In the words of Pericles, Athens"…is open to the world, we never expel a foreigner from learning or seeing". ("Pericles' Funeral Oration", in Thucydides, The History of the Peloponnesian War). It is ATINER's mission to revive the glory of Ancient Athens by inviting the World Academic Community to the city, to learn from each other in an environment of freedom and respect for other people's opinions and beliefs. After all, the free expression of one's opinion formed the basis for the development of democracy, and Athens was its cradle. As it turned out, the Golden Age of Athens was in fact, the Golden Age of the Western Civilization. Education and (Re)searching for the 'truth' are the pillars of any free (democratic) society. This is the reason why Education and Research are the two core words in ATINER's name.

****

The ***Athens Journal of Philosophy (AJPHI)*** is an Open Access quarterly double-blind peer reviewed journal and considers papers from all areas of sports and related sciences. Many of the papers in this journal have been presented at the various conferences sponsored by Philosophy Unit of the Athens Institute for Education and Research (ATINER). All papers are subject to ATINER's Publication Ethical Policy and Statement.

Athens Journal of Philosophy
ISSN NUMBER: 2732-9674
DOI: 10.30958/ajphil
Volume 3, Issue 3, September 2024
Download the entire issue (PDF)

Athens Journal of Philology
Editorial and Reviewers' Board

President's Message

All ATINER's publications including its e-journals are open access without any costs (submission, processing, publishing, open access paid by authors, open access paid by readers etc.) and is independent of presentations at any of the many small events (conferences, symposiums, forums, colloquiums, courses, roundtable discussions) organized by ATINER throughout the year and entail significant costs of participating. The intellectual property rights of the submitting papers remain with the author. Before you submit, please make sure your paper meets the basic academic standards, which includes proper English. Some articles will be selected from the numerous papers that have been presented at the various annual international academic conferences organized by the different divisions and units of the Athens Institute for Education and Research. The plethora of papers presented every year will enable the editorial board of each journal to select the best, and in so doing produce a top-quality academic journal. In addition to papers presented, ATINER will encourage the independent submission of papers to be evaluated for publication.

The current issue is the third of the third volume of the *Athens Journal of Philosophy (AJPHI)*, published by the published by the Philosophy Unit of ATINER

.

Gregory T. Papanikos
President
ATINER

Athens Institute for Education and Research

A World Association of Academics and Researchers

20th Annual International Conference on Philosophy
26-29 May 2025, Athens, Greece

The Philosophy Unit of ATINER organizes its 20th **Annual International Conference on Philosophy, 26-29 May 2025, Athens, Greece** sponsored by the Athens Journal of Humanities & Arts. The aim of the conference is to bring together academics and researchers of all areas of philosophy and other related disciplines. You may participate as stream leader, presenter of one paper, chair a session or observer. Please submit a proposal using the form available (https://www.atiner.gr/2025/FORM-PHI.doc).

Academic Members Responsible for the Conference

- **Dr. Patricia Hanna**, Vice President of Academic Affairs, ATINER, ex-Dean & Professor Emerita, University of Utah, USA.
- **Dr. William O'Meara,** Head, Philosophy Unit, ATINER & Professor, Department of Philosophy and Religion, James Madison University, USA.
- **Dr. Mark Fagiano**, Deputy Head, Philosophy Unit, ATINER & Assistant Professor, Washington State University, USA.

Important Dates

- Abstract Submission: **22 October 2024**
- Acceptance of Abstract: **4 Weeks after Submission**
- Submission of Paper: **28 April 2025**

Social and Educational Program

The Social Program Emphasizes the Educational Aspect of the Academic Meetings of Atiner.

- Greek Night Entertainment (This is the official dinner of the conference)
- Athens Sightseeing: Old and New-An Educational Urban Walk
- Social Dinner
- Mycenae Visit
- Exploration of the Aegean Islands
- Delphi Visit
- Ancient Corinth and Cape Sounion
- More information can be found here: https://www.atiner.gr/social-program

Conference Fees

Conference fees vary from 400€ to 2000€
Details can be found at: https://www.atiner.gr/fees

Athens Institute for Education and Research

A World Association of Academics and Researchers

10th Annual International Symposium on Religion & Theology
26-29 May 2025, Athens, Greece

The Humanities & Education Division of ATINER is organizing its 10th **Annual International Symposium on Religion & Theology, 26-29 May 2025, Athens, Greece**. The aim of the conference is to bring together academics and researchers of Religion, Theology and other related disciplines. You may participate as stream leader, presenter of one paper, chair of a session or observer. Please submit a proposal using the form available (https://www.atiner.gr/2025/FORM-REL.doc).

Academic Member Responsible for the Conference

- **Dr. William O'Meara,** Academic Member, ATINER & Professor, Department of Philosophy and Religion, James Madison University, USA.

Important Dates

- Abstract Submission: **22 October 2024**
- Acceptance of Abstract: **4 Weeks after Submission**
- Submission of Paper: **28 April 2025**

Social and Educational Program

The Social Program Emphasizes the Educational Aspect of the Academic Meetings of Atiner.

- Greek Night Entertainment (This is the official dinner of the conference)
- Athens Sightseeing: Old and New-An Educational Urban Walk
- Social Dinner
- Mycenae Visit
- Exploration of the Aegean Islands
- Delphi Visit
- Ancient Corinth and Cape Sounion
- More information can be found here: https://www.atiner.gr/social-program

Conference Fees

Conference fees vary from 400€ to 2000€
Details can be found at: https://www.atiner.gr/fees

Athens Journal of Philosophy - Volume 3, Issue 3, September 2024 – Pages 111-122

Chatbots, Robots, and the Ethics of Automating Psychotherapy

By Eric B. Litwack[*]

Recent developments in artificial intelligence—AI--have caused considerable discussion among both philosophers of technology and psychotherapists. In particular, the question of whether or not new forms of AI will complement or even replace traditional psychotherapists has emerged as a major contemporary debate. This debate is not entirely new, as it has its origins in the Turing Test of 1950, and an early psychotherapy chatbot named Eliza, developed in 1966 at MIT. However, recent developments in AI technology, coupled with long waiting lists and variable access to psychotherapists have raised the question of machine psychotherapists in an urgent manner. Already, there are psychotherapy apps that one can download onto a standard smartphone and use in lieu of a human psychotherapist. In the near future, this simulacrum of a human therapist may be enhanced by the use of android therapists, programed to duplicate the knowledge and behavior of human therapists instantly, appealing to convenience or self-gratification through technology. This raises a host of ethical questions: can such beings be equally effective, and if so, ought we to reason in a consequentialist manner in their favor so as to increase accessibility and reduce costs through technology? Would there be a psychological difference between automated and potentially anthropomorphic therapy and genuine human therapy, if only a subtle one? Even if a chatbot or robot therapist is transparently such, is there an element of emotional manipulation and potential dishonesty in this interaction? Can the security of clients and their data be thus secured? I will argue that key aspects of chatbot psychotherapy present major ethical and clinical challenges in these areas, although transparent forms of it should not be legally banned.

Keywords: *chatbots, technological momentum, automation, psychotherapy, trust, anthropomorphism, data*

Introduction

In this paper, I would like to point to some ethical questions already emerging in the recent development of chatbot psychotherapy. There will be some interdisciplinary content in this, so as to do justice to the full range and depth of challenges that psychotherapists, applied ethicists and philosophers of technology face in this matter. This also reflects my general thinking, as I am both a practicing psychotherapist and a philosopher.

The ethical questions that strike me as especially urgent around this technology relate to professional trust, data privacy and client security, anthropomorphism, and an excessive degree of instant gratification. The broader philosophical and

[*]Honorary Research Fellow, School of Languages and Cultures (Germanic Studies), University of Sheffield & Faculty, Syracuse University London, (Philosophy and Information Studies), UK.

historical question of technological momentum also raises an important question around chatbot therapy, namely: to what extent are we nudging ourselves into an excessively narrow and harmful technological fix which we may find difficult to correct in the future?

Speaking on Independence Day, 1830, the American statesman Edward Everett declared:

> It is the spirit of a free country which animates and gives energy to its labor, which puts the mass in action, gives it motive and intensity, makes it inventive, sends it off in new directions, subdues to its commands all the powers of nature, and enlists in its service an army of machines, that they do all but think and talk. (Cited in Nye 2007, Location 1939, Kindle Edition).

Now the machines may be, on some accounts, virtually thinking, and they certainly are talking.

The ethical and general technological aspects of this topic are thus at once a matter of urgent and perennial concern. This is because interrogations about the risks and values attendant to technology are at least as old as Greek myth and philosophy, and as recent as ChatGPT. It may therefore be of interest to go back even further than the 19th century, and relate two Greek myths.

Two Greek Myths for the 21st Century

A Greek myth that has always especially struck me as prescient is the myth of Icarus and the sun (Graves 2017, Chapter Ninety-Two). Imprisoned by King Minos of Crete, the artist and engineer Daedalus and his son Icarus attempted to escape with the help of what could be seen as an early form of aviation technology: wings of wax. This initially seemed an efficient and clever type of tool use, until Icarus disregarded his father's warning and flew too close to the sun, melting his wax wings. He then drowned in the Aegean Sea.

Also noteworthy from ancient Greek thought is Plato's recounting of The Tale of Theuth the Egyptian inventor of writing. In *Phaedrus*, Plato recounts how Theuth showed his craft to the god Thamus, who then foretold the decline of memory, linked to the move from an oral to a written culture (Plato 1995).

In both of these Greek myths, the ancients warn us of the dangers of *hubris* and the reckless use of technology, as well as the potential damage of over-dependence on technology. They urge us to handle them with care. Today, as AI and robotics advance at nothing short of spectacular speed, we would do well to heed their counsel and consider where we are going and how we are likely to get there.

Here, I will bracket the important topic of machine consciousness in philosophy of mind. However, I will briefly state that there is a very plausible and strong consensus that consciousness or sentience has not, to date, been instantiated in a machine, and let us assume that this is correct. It therefore must follow that current chatbot therapy involves sentient human beings who are seeking the care of non-sentient computer programs, which they will generally understand to be the

case. Although, on some accounts, machine consciousness may emerge during the course of this century, I believe that the highly probable debate and general controversy around this very process will negate the satisfactory resolution of the problems here discussed. In this article, I will limit myself to some aspects of a topic that greatly concerns me both as a philosopher and a practicing psychotherapist: the ethics of the potential use of Chatbots as replacements for psychotherapists; in other words, the automation of mental health care.

The General Automation Scene

We are living in an era in which automation is progressing at remarkable speed. There is much here both for philosophers of technology and psychotherapists to ponder. By way of general historical and social scientific background, a few key observations: One estimate from the Chief Economist of the Bank of England indicates that as many as half of all UK jobs will soon be automated (Elliot 2015, Litwack 2021). A very significant number of CEOs, engineers, and state agencies see the optimal development of automation as one of the key challenges of our time, and they are wise to do so. However it occurs, it has already changed the way we see and engage in the worlds of technology and work.

There is a lively and broad-ranging debate on the full ethical and philosophical implications of rapid and extensive automation. Optimists, such as Silicon Valley entrepreneur and analyst Jerry Kaplan (Kaplan 2015), hold that automation offers great promise for the fulfilment of human aspirations, and that technological unemployment can be offset both by worker retraining in other sectors, and a well-designed unemployment insurance scheme. Pessimists such as Martin Ford (Ford 2021) are concerned that the speed and breadth of contemporary automation may well overwhelm us over the coming decades, should we not be exceptionally careful. He believes that a basic minimal income scheme is called for to offset this trend.

Part of this perspective stems from concerns over unmanageable technological unemployment, as well as a growing sense that we may be dominated or even destroyed by superintelligent machines of our own creation. Furthermore, some pessimists, such as the philosopher and motorcycle mechanic Matthew Crawford, are concerned that the rise of the machines will lead to the acceleration of an already lamentable tendency towards declining mechanical and craft skills in industrial society (Crawford 2015). One might also speak here of a narcissistic need for instant gratification, of not earning what one has by building it. That isn't always the case; consider the benefits of high-tech improvements in areas such as theoretical science, modern medicine, and public health. But with reference to some skilful activities linked to human interaction, practical skills, and understanding, Crawford is right to remind us that faster and easier is not always better.

Current AI Psychotherapy: Technological Momentum and Nudging?

What is the situation with regards to psychotherapy and the current Chatbot controversy? There are already several specific chatbot therapy programs available; some are free—like Woebot—and others have a paywall. One of the first ethical questions that presents itself is: is the specific roll-out of this technology directing or nudging us in a direction that may be harmful or distorting?

By "nudging", I am following Thaler and Sunnstein's account of how technologies can maximise the odds of certain choices, whether good or bad, by their very form or application (Thaler and Sunnstein 2021). This may well be the case with some psychotherapy chatbots**.**

Several chatbot therapy developers claim to use the common therapeutic approach of cognitive-behavioural therapy (CBT), which is already widely preferred by state and private health care providers, in part due to its short duration and lower cost than long-term psychodynamic therapy. The current pattern of chatbot therapy programs already indicates a further narrowing of therapeutic range, as psychodynamic approaches to talk therapy may be harder to program at a high and sophisticated level using even the most sophisticated contemporary AI. What if, as is likely, psychodynamic approaches to psychotherapy are to be considered preferable for some clients in mental distress? General CBT nudging would then be a clear harm.

If we accept an eclectic model of psychotherapeutic technique that allows for a high degree of client choice, this tendency could represent a case of client nudging through technological momentum, to use a notion from the historian of technology, Hughes (1994). This insightful concept holds that the availability of a particular technology, with few if any saleable and infrastructural alternatives, implicitly directs client or consumer choice in a narrowing manner. This could be seen as a large scale and powerful form of nudging.

Hughes offered this notion as a hybrid alternative both to a radical social constructivism and a rigid technological determinism. He means by the former term a perspective on technology according to which there is little if any necessary causation in technology, it being the product of consumer or users' choices and modification by producers in response to these choices. By the latter, he means the idea that technology is the primary, or even the sole driver of history.

The former perspective, he thought, sees technological development as more open-ended than it actually is, and the latter perspective sees it as more strictly caused and controlled than is the case. A comparison from automotive history may prove useful here.

It is likely a little-known fact today that most of the earliest late nineteenth century vehicle engines were either electric or steam-engines; only about twenty percent of American cars were gasoline-based in 1900 (Appleyard 2022, p. 18). So, in a sense, the current move towards replacing the internal combustion engine with electric vehicles is actually a return to a *past* technology, now seen as the better road not travelled. We didn't know about future climate change and dependency on foreign energy supplies in the early twentieth century. Had we

known about these problems, I am inclined to believe that incentives would have been set up so as to nudge the perpetuation of electric (and possibly steam) vehicles, thereby discouraging the further development of internal combustion engine machines. As such, the massive automotive infrastructure established in the early twentieth century around the internal combustion car might never have occurred. This infrastructure has included service centres, oil wells, engine parts, and targeted advertising.

Reading this as technological momentum doesn't entail seeing it as a matter of fully free consumer choice among equally viable alternatives along the lines of social constructivism, nor does it entail seeing it as an inevitable determinism, no matter what choices were made by social organisations and consumers. Rather, it entails that it is not surprising that one technology came to prevail over others, given its support from a growing infrastructure and its interaction with manufacturers and consumers.

A further key factor in the technological momentum of chatbot therapy may prove to be the mental health care crisis that we are now in, in which numerous prospective psychotherapy clients encounter long waiting lists for therapy and/or treatment costs that they cannot afford. Open access chatbots are already available instantly and at no cost, and even the paywall programs will likely prove to be significantly cheaper than a course of treatment with a human therapist. These chatbots are, of course, available for consultation 24-7, which is not conducive to the development of client autonomy and a mature sense of deferred gratification. Both autonomy and the capacity to defer gratification are necessary components of mental health and maturity, in my view. Their ubiquity within the process is more beneficial than the lack of a waiting period to begin psychotherapy. In effect, clients may be thereby nudged and enabled in narcissistic and dependent behaviours during their use of this technology.

To be fair to CBT, it may be eventually proven to the satisfaction of all psychotherapists to be incomparably superior to its psychodynamic rivals, but this is far from being the case today, and I am doubtful that it ever will be. If contemporary chatbot technology nudges us towards CBT, we may discover that its future popularity facilitates less than optimal results for psychotherapy for the widest possible range of clients.

Trust and Psychotherapy Ethics

A further and fundamental ethical question surrounding chatbot therapy is the question of trust. Trust in professional care providers of all types is fundamental to the very exercise of care, as argued by bioethicists such as Pellegrino (1991, pp. 69–92). If we do not have a full fiduciary relationship with the professional with whom we work, the work itself is not likely to succeed due to mutual suspicion and a coldly legalistic framework, he holds. Trust, in Pellegrino's sense, ought to be considered a necessary condition for psychotherapy, and there must be some of it on both sides of the psychotherapeutic dyad.

How can even a modicum of mutual trust be present, when the therapist is known *not to be even a sentient human being*? Trust entails intentionality and consciousness for its full meaning. Worse, still, what if there is an element of sheer deception and manipulation in the provision of chatbot therapy not employing text-messaging as is currently standard, but a realistic humanoid visual figure that may be difficult or even impossible to distinguish from a sentient human being with his or her own memories, experiences, and emotions upon which to draw?

I take this to be the worst-case scenario, in which the client is wilfully deceived as to the ontology or identity of his or her psychotherapist. This would imply robotic perfection, beyond what Mori (1970) has termed the "uncanny valley" in which we have more of a sense of the eerie or uncanny with robots that are only slightly short of perfection, as opposed to clearly mechanical ones, or perfect ones. This well-known phenomenon in the psychology of technology is likely caused by blurring the line between the human and the mechanical, which would be the very point of advancing chatbot therapy.

It is especially important to note that on any reasonable account of professional ethics, not just procedural, but what might be henceforth termed "full ontological disclosure" is required in the therapeutic process. This means that the client has a right to know whether he or she is talking to a human being or to an AI. One could well here draw on Beauchamp and Childress' classic "Four Principles" approach to health care ethics, and in particular its first principle of respect for the personal autonomy of the client (Beauchamp and Childress 2019, Chapter Four). A lack of full ontological disclosure can only be seen as a violation of such required professional respect. The reasonable legal or professional requirement of a watermark for such technologies would be an effective means of preventing this violation of principle through informed consent and an indication of robotic identity in all interactions, like recording in progress signals.

Furthermore, it is worth noting here that maximising access to psychotherapy at any cost, which would seem to be fundamental to any strongly cosequentialist approach to this general topic, would seem to vitiate this fundamental character of client autonomy, as well as being subject to the reasonable objections from the perspective of anti-utilitarian critiques of radical consequentialism (Smart and Williams 1973). A full treatment of this interesting topic lies beyond the parameters of this article.

The current concern about deep fakes underlines the fundamental values of truth and trust, and this applies not just to media ethics, but to a wide range of social and political questions, including the ethics of psychotherapy (Citron 2019). Should we fail to be careful in this matter, the potential for fraud and deception across the board is obvious.

None of this latter possibility is currently fully possible for psychotherapy, but according to some analysts, the current state of chatbot technology may prove to be preparatory to flawless deep fakes and superintelligent AIs that will make our current technology look nothing short of primitive. Should this technology reach new proportions such that we have reason for believing that it is as sentient as we are, or even beyond us, then this controversy will take on new ethical and

psychological dimensions. I believe that we would then find ourselves faced with a disturbing dilemma of our own creation: is the entity before me conscious or not? If it is conscious and intelligent, that is, capable of intentional states, feelings, and any other psychological phenomena that we normally attribute to human beings, we would likely be inclined to attribute genuine moral and social status to it. We would then see it not only as a true psychotherapist, but as a person as well, with all that this attribution implies with reference to trust, empathy, and intentionality.

The Brilliance and Limitations of the Turing Test

The current controversy over whether or not any program has already passed the Turing Test, and to what extent this clever 1950 thought experiment is sufficiently comprehensive, may thus prove to be only the beginning of a problem of psychological verification, as much as one of moral ontology. We may not, because of the problem of other minds, achieve in this matter a reasonable consensus on whether or not we are dealing with a person, let alone a true psychotherapist. And just as some will be satisfied with an operational model of consciousness, others will not be satisfied that the machine's light is really on, no matter how many tasks an advanced machine with impressive chatbot functionality can perform.

In this matter, I am in agreement with Marcus and Davis (Marcus and Davis 2019, Chapter Eight) that for all of its operational cleverness, the Turing Test measures the capacity of a programmable machine to *simulate* consciousness, rather than offering grounds for the certain *ascription* of consciousness to such a machine.

Many of us will likely want to attribute some degree of sentience to advanced chatbot therapists, if only because doing so will justify our confiding our deepest secrets to them, especially if they are even more realistic than contemporary programs. This confirmation bias may lead to delusionality and excessive anthropomorphism. Nothing is truly conscious just because we want it to be. I would take that to be a narcissistic delusion, which is likely to be reinforced by 24-7 availability and reduced or even zero cost. It could distort our social relations generally, as well as the psychotherapeutic process. In effect, we would thus be blurring the distinction between truth and reality in a culture already in the grips of a tendency towards post-truth and advancing deep fake technology (McIntyre 2018, *passim*).

Psychotherapy Should Not Be Excessively Easy

On this supposedly beneficial therapeutic process, perhaps a clinical word, with reference to psychodynamic approaches. Psychotherapy involves real emotional *work*. Not just for the client, but for the therapist. Psychoanalytic approaches stress the importance of the transference and countertransference, of

the enactment of the childhood complexes of the client in his or her behaviour towards the therapist, as well as the therapist's assimilation of perceptions and feelings, in general perspectives on the client. Non-psychoanalytic forms of psychodynamic therapy tend to focus on a more present-oriented notion; the working relationship between client and therapist.

Furthermore, in attachment psychotherapy, there is an underlining of the importance of an attunement between therapist and client, a meeting of two conscious beings in the therapeutic relationship (Holmes 2014, Chapter Seven).

On any psychodynamic account, replicating this relationship is a necessary condition for the full equivalence between human and chatbot therapists. This, at its full extent, would seem to imply a relationship between two conscious beings. If that is not present, we should probably coin a neologism for the overall process. It might be of at least limited help to some, but it would not be complete psychotherapy.

One possible constructive use for chatbots and companion robots is currently being researched. Studies have been done on the use of companion robots with neurodiverse, and in particular, autistic children (Lucaciu 2013). That may well be an effective and morally unproblematic use of robot and AI technologies, with the proviso that the final goal of the work will be a graduation to more comfortable social relations with human beings.

The Automated Therapeutic Seduction

In her 2015 book, *Reclaiming Conversation*, psychologist Sherry Turkle warns us of the potential for delusionality and manipulation in our interactions with advanced and lifelike technologies. We may come, she thinks, to see companion robots such as Paro the seal, a companion robot that looks and even "behaves" almost like a real animal, as akin to living beings worthy of care and capable of giving it (Turkle 2015, p. 349). However, in the end, Turkle believes that we are here being emotionally seduced or manipulated by what is merely a programmed simulacrum. There is indeed great potential for manipulation and delusionality in this, and Turkle is correct to go on to claim that it is all too easy to provide elderly and disabled patients with such a cute machine for reasons both of cost and convenience. This is being done now, mainly in Japan (Piore 2014).

There is a deep psychological concern in this matter, which I believe can be understood in attachment terms. If, following John Bowlby, we hold that the avoidance of security-producing and exploratory contact with our fellow human beings is a form of developmental disorder, then we are unlikely to be swayed by appeals to a technology of this sort being an inevitable wave of the future (Bowlby 1988, *passim*).

Given that the development of client autonomy and a high degree of relational trust is central to the practice of psychotherapy, one must also ask if these qualities can emerge in our interaction with non-sentient robots and AIs in the same way as in our interaction with conscious human professionals. The hard work of psychotherapy has its ups and downs during an individual's course of

therapy. There will be smooth sessions and challenging ones, containing moments of joy, despair, grief, and anger. Sessions will be missed or arrived at late, illness or technical problems will occur periodically. These must be jointly navigated by therapist and client for months or years, as personality changes in the client are facilitated by facing and dealing with the very difficult imperfections of the process. Could such a volatile and subtle background process be programmed into a future sentient AI? Even if so, how much delusionality and emotional manipulation can be justified along the way? All of this in addition to the likely unresolvable problem of attaining a clear and effective consensus on the question of machine consciousness.

In the Shadow of ELIZA

Once more, philosophers and psychotherapists have much to learn from history. There was an early forerunner of the current and possible future situation. As far back as 1966, Joseph Weizenbaum tested the world's first psychotherapy chatbot, the original ELIZA (Weizenbaum 2015).

Weizenbaum's ELIZA was a standard computer console of its era—and therefore not remotely humanoid. Furthermore, its programming could be described as a very simple version of client-centred therapy. And yet, much to Weizenbaum's own concern, his laboratory subjects often rapidly anthropomorphised it, knowing that it had no more consciousness than a toaster. They confided deep personal problems to it, to the point of requesting that Weizenbaum leave the room in order to do so. For this reason, he became an articulate and passionate critic of what he took to be the irresponsible anthropomorphising of computer and AI technology (Schanze 2010). Weizenbaum, one of the world's greatest computer scientists, believed that we should not share our intimate secrets with a computer.

How much more temptation to anthropomorphise is there with the latest chatbots, such as ChatGPT? And if, as is likely, first deep fake avatars and eventually in the room realistic androids or holograms can provide a simulacrum of psychotherapy algorithmically, then we will be reminded of the fact that sometimes yesterday's science fiction becomes tomorrow's science fact. Future life will look increasingly like a Philip K. Dick story, and many of us will not be entirely pleased with such simulacra.

Deep Data Breaches

I have spoken of sharing our intimate secrets. With chatbot therapy, this entails the use and potential sharing of the most intimate data that anyone can generate. There are privacy commitments to be expected, but data breaches do occur regularly, and so does hacking. This both returns us to the question of trust in professionals, and it raises the ethical and legal matters of client autonomy and liability for violations of privacy. Should there be a psychotherapy data breach

through negligence or hacking, who is responsible? The programmers? The company for whom they work? The possibly sentient AI itself?

Similar concerns are raised by the possibility, already alleged (Bharade 2023), that psychotherapy apps have nudged users to suicide, or potentially to criminality.

There is a parallel here to the ongoing debate over ethical and legal responsibility with reference to autonomous vehicles, and although road fatalities are not at stake in chatbot therapy, we should nonetheless agree that intimate secrets should remain between psychotherapists and their clients. Furthermore, if AIs have indeed nudged users to suicide or criminality, these allegations ought to be investigated and if substantiated, prosecuted along lines comparable to the policies applied to human therapists. However, this would seem especially challenging, given the complex levels of potential design here—much like with the ongoing debate over autonomous vehicles and legal responsibility. One possible solution would be to have software engineers and designers sign off legal responsibility to the companies for which they have produced the apps, with final quality and safety control to be done by said companies which would then be liable for harm due to risks deemed excessive or negligent.

Furthermore, will people feel comfortable seeking the help that they require, if they have good grounds for trusting chatbot therapists less than people? If they trust them more, that would seem to imply social anxiety that should be treated by members of human society. In psychoanalytic and attachment terms, a patten of avoidance would thus be enabled by computer technology.

Psychotherapists, are flawed, like all human beings. We are not fool proof systems. However, we cannot be hacked for nefarious purposes, nor is there a tendency among us to sell clients' intimate secrets for profit to insurance companies or, in some jurisdictions, to government agencies. Once massive profit enters the picture, we are at risk of what Shoshana Zuboff has termed 'surveillance capitalism', in which our data is problematically shared for profit. Or it could be used, notably by authoritarian states, in order to control and punish their citizens. This, too, is already well underway, with the potential for totalitarianism beyond the imagination of history's worst dictators. One might add the risk of extortion in cases in which unscrupulous hackers could threaten to expose psychotherapy clients' intimate secrets if they do not send funds immediately.

I do realise that this may seem to be dystopian doomsaying, but I am here inclined to agree with Ludwig Wittgenstein that:

> the work of the philosopher consists in assembling reminders for a particular purpose (Wittgenstein 1976, Part I, Section 126).

If we hold this to be true, at least when controversial matters are involved, then we not only can risk playing Cassandra, we must do so. So much is at stake here that moral responsibility entails caution of the sort that has motivated the recent call for a moratorium of at least six months—not a cessation--on chatbot development so as to explore optimal governance procedures that would include legislation and advanced security measures (Future of Life Institute 2023). Leading computer scientists and Silicon Valley entrepreneurs are among those

who have signed this petition, which acknowledges both the promise of AI and its potential perils for the future of humanity.

With reference to transparently chatbot psychotherapy, we would do well to remain in this spirit. Let us acknowledge its potentially beneficial aspects, as well as its indicated pitfalls, while advocating greatly improved governance and ethical judgment. Low or even no cost therapy, and shorter (but not necessarily no) waiting time to begin a course of therapy are indeed advantageous developments, even if we are talking about a stop gap measure that is in need of empirical study concerning outcomes.

With reference to legislation, an all-out ban on the *transparently* virtual varieties of psychotherapy would be excessively draconian within a democratic society. We have to allow for a reasonable range of choices that may be far less than optimal, as long as harm is contained, and reducible**.**

That is likely the case here, given the foreknowledge of the psychotherapy clients of what they are talking to—full ontological disclosure--and the potential for the wide dissemination of critical information encouraging richer and more authentically human forms of this work. Sometimes we must look to the state to prohibit severe harm, but the object of our discussion here may be a more moderate harm, or settling for a less than optimal activity as a technological quick fix. In that sense, transparently virtual forms of psychotherapy may be, at best, comparable to vitamin supplementation in lieu of a well-balanced diet. Not as good, but better than severe malnutrition, thanks to advanced technology. Should more than anecdotal evidence emerge indicating that this form of the technology is worse than that, then state regulation should certainly be considered. The trick then would be to counteract the potentially huge underground economy that would accompany the prohibition or severe curtailment of the technology. It may well not be worth an attempt at strict prohibition.

Having said this, it is my belief that much more must be done to provide access to psychodynamic therapy of various schools with human beings at reasonable cost and in reasonable time. This may not be easy. But the ethical and psychological aspects of this problem necessitate a valiant effort. Nudging or even requiring psychotherapists to provide some concessionary rates will likely be part of this strategy, as will the training of more professional therapists.

Above all else, we should be cognisant of the need to distinguish between sharing intimate secrets with a trained and conscious human being who has a genuine professional and personal concern for us, with what (for now) is in the end an obvious simulacrum with no more consciousness than a toaster.

References

Appleyard J (2022) *The Car: The Rise and Fall of the Machine that Made the Modern World*. London: Weidenfeld & Nicolson.

Beauchamp T, Childress J (2019) *Principles of Biomedical Ethics*. New York: Oxford University Press.

Bharade J (2023) *A Widow is Accusing an AI Chatbot of Being a Reason Why her Husband Killed Himself*. Business Insider. Available at: https://www.businessinsider.com/widow-accuses-ai-chatbot-reason-husband-kill-himself-2023-4.

Bowlby, J. (1988) *A Secure Base*. Abingdon: Routledge.
Citron D (2019) *How Deep Fakes Undermine Truth and Threaten Democracy*. TEDSummit. Available at: https://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy.
Crawford M (2009) *Shop Class as Soulcraft: An Inquiry into the Value of Work*. London: Penguin Press.
Elliot L (2015) *Robots Threaten 15m UK Jobs, Says Bank of England's Chief Economist*. Available at: www.theguardian.com/business/2015/nov/12/robots-threaten-low-paid-jobs-says-bank-of-england-chief-economist.
Ford M (2021) *Rule of the Robots: How Artificial Intelligence Will Transform Everything*. London: Basic Books.
Future of Life Institute (2023) *Pause Giant AI Experiments: An Open Letter*. Available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.
Graves, R. (2017) The Greek Myths. London: Penguin Press.
Holmes J (2014) *John Bowlby and Attachment Theory*. Hove: Routledge.
Hughes TP (1994) Technological Momentum. In L Marx, MR Smith (eds.), *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge: The MIT Press.
Kaplan J (2015) *Humans Need Not Apply*. New Haven: Yale University Press.
Litwack EB (2021) Bertrand Russell on Automation: How Necessary are Human Beings? *Bulletin of the Bertrand Russell Society* 163: 61-69. Available at: https://bertrandrussellsocietyorg.files.wordpress.com/2022/01/32e20-brs-bulletin-spring-2021.pdf.
Lucaciu I (2013) Robots: *The Answer for Treating Children with Autism Spectrum Disorder?* Available at: http://www.theneuroethicsblog.com/2013/07/robots-answer-for-treating-children.html.
Marcus G, Davis E (2019) *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon Books.
Marx L, Smith MR (1994) *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge: The MIT Press.
McIntyre L (2018) *Post-Truth*. Cambridge: The MIT Press.
Mori M (2012) *The Uncanny Valley*. Available at: https://spectrum.ieee.org/the-uncanny-valley.
Nye D (2007) *Technology Matters*. Cambridge: The MIT Press.
Pellegrino ER, Veatch RM, Langan JP (1991) *Ethics, Trust, and the Professions*. Washington DC: Georgetown University Press.
Piore A (2014) Will Your Next Best Friend Be a Robot? *Popular Mechanics*. Available at: https://www.popsci.com/article/technology/will-your-next-best-friend-be-robot/.
Plato (1995) *Phaedrus*. Indianapolis: Hackett Publishing Company.
Schanze J (2010) *Plug & Prey*. Maschafilm. Available at: http://www.plugandpray-film.de/en/.
Smart JJC, Williams B (1973) *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.
Thaler R, Sunstein C (2021) *Nudge: The Final Edition*. London: Allen Lane.
Turkle S (2015) *Reclaiming Conversation: The Power of Talk in a Digital Age*. New York: Penguin Books.
Weizenbaum J (2015) *Islands in the Cyberstream: Seeking Havens of Reason in a Programmed Society*. Sacramento: Litwin Books.
Wittgenstein L (1976) *Philosophical Investigations*. Oxford: Basil Blackwell.
Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

Athens Journal of Philosophy - Volume 3, Issue 3, September 2024 – Pages 123-134

Aquinas's Claim: Love of Neighbor as Oneself is Self-evident

By William O'Meara[*]

Thomas Aquinas has affirmed that love of neighbor as oneself is self-evident. We shall distinguish for Aquinas three forms of self-evidence: (a) evident in itself but not to people, (b) self-evident in itself and to all human thinkers, and (c) self-evident in itself but only to expert thinkers. We shall consider only (b) and (c). (b) We shall develop an analysis attempting to show that love of neighbor as oneself is self-evident in itself and to all human thinkers, however, in evaluation we shall argue that the principle is not self-evident to all human beings since not all people love themselves wisely. Further, we shall consider arguments from Kant that some maxims of the will are self-contradictory to affirm and hence immoral whereas the denials of other moral commands are not self-contradictory to deny. (c) However, we shall develop an argument affirming that love of neighbor as oneself is self-evident to expert thinkers, namely, self-evident to some human beings who have made the wise decision to live by the Aristotelian Principle. Finally, we will consider two objections to the self-evidence of love of neighbor as oneself and answer them, noting the importance of wise love of others as one wisely loves oneself. Such wise loving of oneself and others avoids simplistic misinterpretations of love of neighbor as oneself.

Keywords: *Thomas Aquinas, self-evidence, Aristotelian principle, love of neighbor as oneself, morals*

Thomas Aquinas writes in the *Summa Theologiae* that love of neighbor as oneself is self-evident and that the commandments from "Honor thy father and mother" to "Thou shalt not covet thy neighbor's goods" are self-evident as basic conclusions of the command to love one's neighbor as oneself (Aquinas. I, II, Q. 100).

Aquinas identifies three forms of self-evidence: (a) A statement can be self-evident in itself but not to humans such as the statement that God necessarily exists. God knows with certitude that the predicate, necessary existence is included in the concept of the subject, God as the Perfect Being. However, since humans have identified many corporeal things such as animals as Divine, humans do not have the concept of God clearly in their minds; and so they do not know that God necessarily exists. (b) A statement can be self-evident in itself and also to humans such as the statement that a being cannot both be and not be at the same time and in the same respect, or more simply, being cannot be non-being. Knowledge of the principle of non-contradiction is self-evident to all thinking human beings. (c) A statement can be self-evident in itself such as the Pythagorean Theorem, that the square upon the hypotenuse of a right triangle is equal to the sum of the squares of

[*]Professor, James Madison University, USA.

the two legs of that triangle, but only known as such by experts with advanced knowledge. Only those with knowledge of geometry know the self-evidence of the Pythagorean Principle (St. Thomas Aquinas, *Summa Theologiae*, Part One, q. 2, a. 2.). We shall examine whether love of neighbor as oneself is self-evident as (b) or as (c).

Is Love of Neighbor as Oneself Self-Evident to all Humans?

Aquinas is affirming that love of neighbor as oneself is self-evident in the format of the second kind, that is, (b) self-evident in itself and also to all thinking human beings. Let us assume that we can formulate love of neighbor as oneself in both a negative way and in a positive way. We shall turn to very early religious formulations of the love of neighbor as oneself since Aquinas himself is commenting on the religious background of love of neighbor. A formulation in a negative way was stated by Rabbi Hillel, "That Which Is Hateful to You, Do Not Do to Your Fellow! That is the Whole Torah; The Rest is Interpretation" [Babylonian Talmud, Shabbat 31a]. A formulation in a positive way was stated by Jesus, "So whatever you wish that others would do to you, do also to them, for this is the Law and the Prophets (Matthew 7:12)." In the light of these two formulations, negative and positive, of the Golden Rule, we shall be reflecting in this paper upon whether both the negative and positive versions of love of neighbor as oneself are self-evident in the consciousness of all human beings. We will not engage in argumentation about whether the negative formulation or the positive formulation is better but only reflect on whether or not each version is self-evident to all thinking humans.

Here is a first analysis of whether the negative and positive versions of the Golden Rule, "Do not do to others what is hateful to yourself," and "Do unto others as you would have them do onto you," are self-evident. It is self-evident that if one wishes to love one's own self by not doing things that harm one's own life and well-being and by positively doing things that affirm and develop one's own life and well-being, then one should not in those efforts do things that harm others. For if one would allow own's own self to harm others, then those others could reasonably be willing to harm that original self, and thus harming others would have the consequence of thwarting the protection and overall development of oneself. So just as one should not harm oneself directly and should advance one's own development, so also one should not harm others and should assist in the proper development of others, for example, when children and teenagers need help in learning basic mathematics and also in learning advanced mathematics such as algebra, geometry, trigonometry and calculus.

In effect, this proposed defense of the self-evidence of love of neighbor as oneself is similar to Kant's defense of his first formulation of the Categorical Imperative, the Universalization Principle, "So act that the maxim of your will may be a universal law for all rational beings (Kant, Part II)." Kant's defense of this principle is that it is self-contradictory to give yourself one maxim, "You may hurt others," and also a contradictory maxim to others, "Others may not hurt

me." Here is how Kant's argument would proceed. As a rational being, reason obliges a person, argues Kant, to act in a self-consistent way. If I say to myself that I will act in a specific way towards others although I would not want others to act that way towards me, Kant would say that such a maxim of the will is not self-consistent, but is irrational and immoral. For I am treating myself as a rational being in a way different than I am treating other rational beings. If I ask what is wrong with being inconsistent in my treatment of myself and others, Kant would reply that there is something in every human being that makes him resist and resent being treated as a thing instead of a person. Just as I as a reflective and therefore free being resent being treated as unreflective and unfree, so also do all other persons as reflective and free beings resent being treated as things. When I claim for myself a dignity as a reflective and free being I am thereby asserting a claim to dignity. Therefore, inconsistency in the way I would treat myself and others in effect destroys any rational claim I make for myself as someone valuable for my own sake. If I can deny to others their value for their own sake in the maxims of my actions, then I am logically denying to myself any right to claim value for my own self as rational and free. Consequently, the best formulation of the basic value or moral law is the formulation of the Second Categorical Imperative: "So act as to treat humanity, whether in thine own person or in that of any other, in every case as an end, never as means only (Kant, Part II)."

So, Kant has claimed that that the maxims, "I may hurt others, and others may not hurt me," are self-contradictory. But a closer analysis here would reveal that instead these maxims, "I may hurt others," and "I may not hurt others," constitute a true self-contradiction and actually paralyze oneself from any action. However, the maxims, "I may hurt others, and others may not hurt me," would lead to conflict between self and others, but not involve the self in being self-contradictory. For a person could say rationally without self-contradiction that the self who is a robber baron may hurt others because the robber baron believes he would be victorious in any battle between the self as a robber baron and the other who are the victims of the robber baron. It is not self-evident that the robber baron is self-contradictory. John Stuart Mill offers a similar evaluation of Kant's first formulation of the Categorical Imperative, writing that when Kant:

> begins to deduce from this precept [the Categorical Imperative] any of the actual duties of morality, he fails, almost grotesquely, to show that there would be any contradiction, any logical (not to say physical) impossibility, in the adoption by all rational beings of the most outrageously immoral rules of conduct. All he shows is that the *consequences* of their universal adoption would be such as no one would choose to incur (Mill, *Utilitarianism*, ch. 1).

The maxims, "I may hurt others" and "Others may not hurt me," are not logically self-contradictory and also not a physical impossibility. The adoption of such contrary rules can lead to physical difficulties, what Mill calls consequences

that people would strive to avoid. These consequences are not a physical impossibility, but they are dreadful, leading to horrible conflicts between people[1].

We noted above that Kant's defense of the first formulation of the Categorical Imperative is dependent upon his second formulation since the reason for not harming others is precisely that it is self-contradictory for the self to claim value for oneself and not for all other persons. Here is a defense of why we must value all others as we should value ourselves. If we grant that we each value ourselves, why should we value all other selves? Kant affirms that the only way in which I can fully appreciate my own value for myself is if I also value all other selves if the conception of the value of self is to have its full effect upon me. As Kant has written,

> The natural end which all men have is their own happiness. Now humanity might indeed subsist, although no one should contribute anything to the happiness of others, provided he did not intentionally withdraw anything from it, but after all this would only harmonize negatively not positively with humanity as an end in itself, if everyone does not also endeavour, as far as I him lies, to forward the ends of others. For the ends of any subject which is an end in himself ought as far as possible to be my ends also, if that conception is to have its full effect with me (Kant, *Metaphysic of Morals*, Part II).

Well, we could grant that point of Kant that we first can learn of the value of ourselves for our own self in some intimate family community in which we have a richer effect of the value of our shared humanity in our small community, but why should I value all the selves in another country? Of course, what would happen if the people in another country valued only themselves but not the people in my country? It seems highly probable that this narrow valuing of selves only in my own family or nation could lead to terribly destructive effects that could come back to harm myself, my family, and my nation because other people in other countries would not respect my value as a person and neither the value of my family members nor the value of my fellow citizens in my country (Kant, *Metaphysic of Morals*, Part II).

In evaluation of this proposed defense of Kant's second Categorical Imperative, we may note that we only affirmed the probability of terrible effects coming back to the self or country that only valued itself. It is not self-evident that one who values oneself must also value others in a similar way.

Here is how Kant could reply to the analysis that respect for the value of all other persons is not necessarily included in the concept of respect for the value of oneself in six steps:

1. If I value myself as a male, then logically I should value all other males.
2. If I value myself as of a specific ethnic descent, then logically I should value all others of a similar ethnic descent.

[1]This paper has been deeply influenced by Mill's evaluation of Kant and is essentially a development of Mill's analysis of Kant's argument as failing to establish any self-contradiction in outrageously immoral rules of conduct.

3. But it is not my gender or ethnic descent that is at the heart of my being, rather it is my humanity, my consciousness in reasoning and making free choices that makes me a human being. (Valuing myself primarily for my gender or for my ethnic descent would be a valuation based on mere accidents or non-essential characteristics, whereas valuing myself for my rationality and freedom would be a valuation based on my essential characteristics without which I could not function as a human person. To say this point in another way, gender and ethnic descent are neither necessary nor sufficient conditions for being self-conscious as a human being, but rationality and freedom are both necessary and sufficient conditions for being self-conscious. It is important to recognize that one needs to make a commitment to one's abilities of rationality and freedom as central to one's identity and full development; in doing so one is living by what John Rawls has called the Aristotelian Principle which we analyze later in this paper (Rawls, p. 427).)
4. So, if I value myself as a human being, then logically I should value all human beings (and any other mammals who have self-consciousness, rationality, and freedom).
5. I necessarily value myself, my consciousness, rationality, and freedom, whenever I consciously choose to value anything, for example, like basketball or tennis because conscious valuing which is rational and free is the necessary condition for making any choices.
6. So, I should logically value all human beings and other beings with precisely such attributes of consciousness, rationality, and freedom just as I necessarily value my own consciousness, rationality, and freedom.

Over against this line of argument on behalf of Kant for the self-evidence of respect for others being included in the concept of self-respect, a contrary argument may be made. Is it really necessary that every self must value itself? Could we not find evidence in the fact that there are people who do not respect themselves at all? For example, Rachel Bachner-Melman, a clinical psychologist at Hadassah University Medical Center in Jerusalem who specializes in eating disorders, has seen the impact of extreme selflessness on the anorexic young women who populate her ward. "They are terribly sensitive to the needs of those around them," she said in an interview. "They know who needs to be pushed in a wheelchair, who needs a word of encouragement, who needs to be fed." Yet the spectral empaths will express no desires of their own. "They try to hide their needs or deny their needs or pretend their needs don't exist," Dr. Bachner Melman went on. "They barely feel they have the right to exist themselves." They apologize for themselves, for the hated, hollow self, by giving, ceaselessly giving (Angier, "The Pathological Altruist . . .,"). These spectral empaths are, admittedly, out of the realm of normal behavior. Consequently, if we were to omit consideration of such people and if we were to defend Kant's universal valuation of rationality and freedom for all normal selves, perhaps Kant's argument might be accepted as based on a hypothetical imperative in the following way: If a person loves oneself, that is, respects one's rationality and freedom as having inner dignity, then

logically one should love all other selves, that is, respect all others in their rationality and freedom as having dignity. If I were to attempt to live out the choice that I have value but others do not, these others would not logically have to adopt an attitude of valuing me. So, avoiding this negative consequence of others not valuing me would be a good pragmatic reason for me to value others as I value myself.

The problem, however, still remains that mere consciousness of one's own rationality and freedom is insufficient to establish in these people who are outside of normal behavior the deep valuation of their own rationality and freedom and that of others. So, Kant's argument from rationality and freedom of the self does not establish with certitude that the valuation of others is founded necessarily in the concept of self-respect of one's own value. Yes, many people may combine such valuation of self with a similar valuation of others, but it is not logically necessary that one must value others as one values oneself.

In accord with Kant's attempt to show that self-valuation has a necessary connection with valuation of others for their own sake, Kant offers a solution to this following ethical problem that he raised for himself:

> A fourth [person], who is in prosperity, while he sees that others have to contend with great wretchedness and that he could help them, thinks: "What concern is it of mine? Let everyone be as happy as Heaven pleases, or as he can make himself; I will take nothing from him nor even envy him, only I do not wish to contribute anything to his welfare or to his assistance in distress!" (Kant, *Metaphysic of Morals*, Part II).

Kant responds to this problem, affirming that we have a moral duty to help others in great need because there is a contradiction between saying that a person need not now help others and saying later that when the self, as a person, is in great need, that others ought to help that person (Kant, *Metaphysic of Morals*, Part II).

However, Kant's solution of this problem has the following internal flaw. For Kant acknowledges that a person could make a universal maxim that everyone must solve their own problems, and there is no inner contradiction in that universalization. However, Kant believes that one would regret such a maxim most probably in the future when one is drowning or one's house is threatened by a fire because then one could not solve the problem by one's own efforts and would want to have the help of other. Here Kant in appealing to the bad effects of such a maxim in relying upon consequences, as John Stuart Mill has pointed out, whereas he wants to hold that that universal maxims are to be held because they are inherently not contradictory, not because of bad consequences (Kant, *Metaphysic of Morals*, Part II).

Kant's solution of this problem is not consistent with his concept of humanity as an end in itself. Kant acknowledges that a person might refuse to help such others and that this refusal does not violate humanity in one's own person as an end in itself. However, Kant says that action should not only not violate humanity as an end in itself; action should also advance this end. Hence, it is immoral not to help others when they are in immediate and great need and cannot save or protect themselves. This solution by Kant violates his own guideline that moral decisions

should not be based on the consequences of action.[1] Kant had said that an act should not be judged in the light of its consequences or ends to be effected. Kant had said that the principle of humanity should not be used as a positive goal but as a negative principle which prohibits actions against the value of the human being because the inherent valuing of oneself rejects any actions which would harm other selves. Kant has failed to show that in acting for one's own benefit that one must act to help others in great need. We can imagine that one person might join with 99 others to work together to help each other when in great need but there is no inner contradiction in their joining of this group of 99 other people that would require them to act to prevent great harm to others outside their own special group of 99 people.

In summary, so far, we have presented analyses rooted in Aquinas and Kant that love of neighbor as oneself is self-evident to all thinking human beings, but we have argued against Aquinas and Kant that these analyses have failed to show this self-evidence. We can now take up the question of whether the love of neighbor as oneself is self-evident to the expert thinker just as the Pythagorean Theorem is self-evident to the expert geometrician.

Is Love of Neighbor as Oneself Self-evident to the Expert Thinker?

We shall reflect on what John Rawls has identified as the Aristotelian Principle to see if all who live by that principle see the self-evidence of love of neighbor being included in the concept of self-love[1].

John Rawls has identified a principle implicit in Aristotle's analysis of human nature and behavior which he calls "The Aristotelian Principle" (Rawls, p. 427). This principle of deep motivation affirms that "other things equal, human beings enjoy the exercise of their realized capacities (their innate or trained abilities), and [that] this enjoyment increases the more the capacity is realized, or the greater its complexity" (Rawls, p. 427). Although it is not necessarily true for all people, generally we may say that most people will prefer to express in action those abilities which are more challenging to realize and that people will enjoy that expression even more, the more difficult that those activities are to realize. For example, a biologist will enjoy doing advanced research on cloning more than simply repeating the experiments of Pasteur which established that spontaneous generation of life from rotting organic matter could not occur. Rawls further notes the importance of self-respect as a primary good for the self in the expression of the Aristotelian Principle in our behavior. To understand this point, we may offer an analysis of someone who is a wastrel, a good-for-nothing, as contrasted with someone who exemplifies the Aristotelian Principle. A wastrel has no challenging goal to realize, such as contributing to the advancement of biology and also has no effective work habits of having learned inorganic chemistry, organic chemistry and fundamental field of biology and many other disciplines of advanced biology which would enable a person to be able to make a lasting contribution to this

[1]The analysis that follows is similar to my analysis in O'Meara (2015). The application to the self-evidence of love of neighbor as oneself is completely new.

science. In contrast, one who lives by the Aristotelian Principle in the study of biology will have a challenging goal, perhaps that of contributing to the advancement of human knowledge, and will also have developed the difficult means of practicing the mastery of biology. In contrast with the good-for-nothing who has no challenging goal and no challenging means to that goal, the biologist has both a challenging end and a continuing mastery of the difficult means appropriate for achieving this end (Rawls, p. 440). Most importantly, this biologist's self-respect, founded in both a challenging end and a challenging means, occurs best in a community of fellow biologists who themselves can appreciate both the true difficulty of a great contribution to biology and the difficult means of developing the mastery of biology hitherto which sets the groundwork for the possibility for a new contribution to the field. The fact that other biologists value both one's own goal of contributing to the advancement of the biological sciences and the difficult means of striving towards that goal helps to contribute to one's own valuation of the worthwhileness of one's goal and of the difficult means. No person's estimate of the person's own self-respect is an island unto itself, but people can be all intertwined in this community, for example, of biologists, and their mutual appreciation of each other's contributions reinforces our own appreciation and valuation of our selves. In summary, the conditions for persons respecting themselves and one another require that the fundamental goals of our actions be well thought out and complementary to each other so that, for example, the contributions of various members of the community can fit together in such a way that one scholar advances knowledge of cell division, another brain chemistry, a third new medicines for treating various cancers, and a fourth who teaches beginning students of chemistry and biology to high school students. These goals would fit together well and not thwart various goals of other scholars. Each one in the community of scholars both would appreciate the worthwhileness of the various goals sought in the field of study and also would appreciate the challenging means practiced by all as required to achieve their valuable goals (Rawls, p. 441).

Assuming that a person has made the inner decision to live by the Aristotelian Principle, should we say that this person's decision to love oneself, for example, as a biologist, who strives not only to avoid harm to the science of biology, but also to advance the science of biology through new discoveries, necessarily involve love of other biologists, essentially in the form of respect for them and, indeed, of anyone who strives to know the truth? This love for others is not an emotional love, but the love of respect in which one does not do harm to others and in which one strives to be of assistance to others when it is reasonably appropriate. For example, should one biologist not do damage to another biologist's experimental data, and should one biologist also point out to another that a mathematical error has been committed by the second biologist if such is the case? The answer to these two questions is: Yes, definitely yes! For respect for oneself as a biological scientist necessarily involves respect for all scientists and all who seek to know the truth. As we have affirmed in our summary for Rawls above: "No person's estimate of the person's own self-respect is an island unto itself, but people can be all intertwined in this community, for example, of biologists, and their mutual

appreciation of each other's contributions reinforces our own appreciation and valuation of ourselves."

In objection to this claim that the commitment to the Aristotelian Principle necessarily links valuation of the self as an end in itself with the valuation of all others who can be ends in themselves, we may develop another example. Imagine that a person has made a commitment to the Aristotelian Principle by choosing the difficult goal of becoming a master of the criminal arts by a careful development of the difficult means of learning all kinds of accounting fraud and of avoidance of justly due taxes. In this imagined case, one's valuation of oneself as a criminal would necessarily involve respect for other criminals, creating honor amongst thieves, in one's criminal syndicate, but there would be no valuation of others who would be the victims of their criminal activity. Consequently, the Aristotelian Principle would not appear to show that love of all others is necessarily involved in love of self.

However, here the argument of Kant could be offered that one who would commit to joining such a criminal syndicate would later regret such a choice when that person has been caught and offered the chance to turn state's evidence against the leaders of a criminal syndicate. If Kant's argument against one who would adopt the principle of not helping others who are in great distress as involving a contradiction with what a person would will later when that same person would be is desperate need, then we may also argue as follows about the community of those who live by the Aristotelian Principle that the only consistent choice without contradiction, both now and later, for affirming the value of oneself as an end in itself is to join the community of all rational beings who also affirm the value of all rational and free beings as ends in themselves. So, if we accept Kant's argument about adopting now and in the future consistent maxims of the will for all and any who might be in desperate need, we may conclude that for the person who has made the deep commitment to live by the Aristotelian Principle, love of oneself, that is, respect for oneself, necessarily involves love of neighbor, respect and appropriate assistance when reasonable to another, as the only consistent choice which will not be contradicted by a later choice, regretting one's original but inadequate choice of love of oneself. Consequently, love of neighbor, respect for and reasonable assistance to neighbor as oneself, is self-evident, not to all humans, but to the expert thinkers who live by the Aristotelian Principle. The great insight of Aristotle that the ethical mean lies between the two extremes of too much and too little needs to be kept in mind. A person could have too little love of and respect for oneself or too much. So also, a person could have too little love of and respect for oneself, as the spectral empaths do, or even too much as person might do when suffering from hyper-empathy disorder [Hyper Empathy Disorder, internet]. The person who has achieved or come close to the Golden Mean of proper love of and respect for oneself will necessarily understand that the Aristotelian Principle requires one to live with wise love of and respect for others. It is important to note that wise self-love and wise self-love of others cannot be precisely defined with an abstract formula. Wise love needs a practical learning in which one learns from mistakes in loving oneself too much as the social deviant narcissist does and in loving oneself too little as the pathologically dependent

spectral empath does. So, a person has to really interact with others and receive responses that encourage the self not to love oneself excessively and not to love oneself defectively.

Two Objections to the Moral Principle of Love of Neighbor as Oneself

A first objection is that we should love others, not as we love ourselves, but as the others wish to be loved. For example, if I prefer chocolate ice cream for dessert, it does not follow with necessity that I should serve only chocolate ice cream to my guests. This objection may be easily answered because the principle of love of neighbor as oneself should not be applied blindly as a mathematical formula to all people and circumstances but with wisdom. Wise reflection upon the needs of oneself and of others reveals that needs and wants can vary from person to person and that, consequently, what may be satisfactory to oneself may not be so for the other. Negative rules are easier to generate. For example, as the self would not want to risk suffering from placing one's own hand in a fire, neither should we risk that others would burn their hands in a fire. In contrast to negative rules, positive rules are harder to generate since we may not be able to simply use our own feelings or wants as the criterion, but we must be sensitive to the feelings and wants of others. One's preference for dessert may be one thing, but other people may have their own significant preferences. So, we need to understand that we should love, that is, respect and assist when it is reasonable, others wisely as we love, that is, respect and develop, ourselves wisely.

A second objection is that a judge in a court should not sentence a convicted felon to prison since the judge would not want to be imprisoned either. Consequently, morality and law and order would fall apart if we follow love of neighbor as oneself, the challenge is. Again, here the answer is that this principle should be understood as love of neighbor wisely as one loves oneself wisely. Wise reflection discloses that adult morality requires that we hold adults to full responsibility for their moral misdeeds and that, consequently, the wise judge needs to imprison the convicted felon in accord with the punishments called for by the law. The wise judge would recognize that if the judge commits a felony, the judge should reasonably will that the judge should be punished in accord with the law upon conviction for the felony. The principle, in summary, is that we should love, that is, respect and assist when it is reasonable, others wisely just as we should love, that is, respect and develop, ourselves wisely.

Summary

Thomas Aquinas has affirmed that love of neighbor as oneself is self-evident. We have argued that the principle is not (b) self-evident to all human beings since not all people love themselves wisely. For some people such as spectral empaths who care so much for others in anorexia wards in hospitals, nevertheless have lost their own fundamental decision to care for themselves wisely. Also, the principle is not self-evident to social deviants who tragically may

never have developed the ability or perhaps who did develop this ability but have subsequently lost the ability to care for others for their own sake However, we have developed an argument affirming that love of neighbor as oneself is (c) self-evident to some human beings who have made the wise decision to live by the Aristotelian Principle who seek self-excellence in their difficult goals and in their difficult means to those goals, thereby seeking to live in a community of respect for those seeking challenging goals through challenging means. So, we concluded, if Kant's argument about adopting now and in the future consistent maxims of the will for offering aid to all and any who might be in desperate need, that for the person who has made the deep commitment to live by the Aristotelian Principle, love of and respect for oneself then necessarily involve love of and respect for neighbor as the only consistent choice which will not be contradicted by a later choice, regretting one's original but inadequate choice of love of oneself. Finally, we considered two objections to the self-evidence of love of neighbor as oneself and answered them, especially the case of a judge sentencing a convicted felon, noting the importance of wise love of and respect for others as one wisely loves and respects oneself. Such wise loving of and respect for oneself and others avoids simplistic misinterpretations of love of and respect for neighbor as oneself.

References

Angier N (2011, October 3) *The Pathological Altruist Gives Till Someone Hurts*. The New York Times. Available at: https://nyti.ms/2NE5XzK.

Aquinas T (n.d.) *Summa Theologiae*. Available at: newadvent.org/summa.

Hillel R (n.d.) *Babylonian Talmud.* Shabbat 31a. Available at: ejiltalk.org/that-which-is-hateful-to-you-do-not-do-to-your-fellow-that-is-the-whole-torah-the-rest-is-interpretation- from-the-elder-hillel-in- babylonian-talmud-shabbat-31a.

Hyper Empathy Disorder. Available at: https://www.theedgetreatment.com/hyper-empathy-disorder-what-happens-when-you-feel-too-much/.

Kant I (n.d.) *The Project Gutenberg EBook of Fundamental Principles of the Metaphysic of Morals*. Available at: https://www.gutenberg.org/files/5682/5682-h/5682-h.htm#.

Mill JS (n.d.) *The Project Gutenberg EBook Utilitarianism.* Available at: https://gutenberg.org/cache/epub/11224/pg11224-images.html.

O'Meara W (2015) The Aristotelian Principle in Mill and Kant. *Athens Journal of the Arts and Humanities* 2(1): 9–18.

Rawls J (1971) *A Theory of Justice.* Boston: Harvard University Press.

Athens Journal of Philosophy - Volume 3, Issue 3, September 2024 – Pages 135-148

Nature of Vedic Ethics and its Critique as Soteriology

By Swagata Ghosh[*]

The present paper deals with the idea of understanding Vedic ethics as a code of righteous living, in the light of Mīmāṁsā philosophy. The paper also intends to reflect upon the possibility of such methods as a means of attaining liberation. In other words, the Vedas provide us with prescriptive codes of right and wrong actions. It commands us about duties and non-duties, through the performance of rituals, in order to lead a good life. We know that human endeavours are primarily based on attaining the desired, and to prevent the unwanted ends. Hence, the entire effort of human actions lies in the fact that we want to attain the cessation of sufferings. This paves the way for studies in Soteriology, and the question arises that, could the Vedas be considered as a literature on Soteriology. The paper consists of a number of sections. The initial ones involve the linguistic analysis of the Vedic statements based on Sanskrit grammar and semantics. These are aimed at depicting the manner in which Vedic sentences act as prescriptive ethical codes. The latter ones deal with the questions raised by the opposing schools, namely Sāṁkhya-Yoga, against the idea of considering Vedas as a supreme sanction of means, leading beings to their salvation from empirical sufferings, and their plausible responses.

Keywords: *duties, ends, sacrifices, knowledge, liberation*

Introduction

Indian philosophical conception of morality is widely based on the Vedic notions of right and wrong. The philosophical traditions not only restrict themselves in advancing various theories on ethics regarding the standard of evaluation of voluntary acts, or trying to determine the import of ethical terms, or even factors influencing moral judgement and the like, rather the orthodox traditions hold that the ultimate sanction of morality is the Vedas. This is because of the fact that the various schools of thought in the Indian tradition are primarily based on Soteriology. Thus, we find that any discussion on morality, however, is two-fold – one aspect deals with the qualitative evaluation of intentional actions, while the other prescribes or commands ways of righteous living, which eventually lead to the salvation of individuals. Both the functions are deeply related to linguistic employments, and hence, our main contention in this paper would be to carry out an analytic discussion on morality as expressed through linguistic usage in the Vedas. Further, the idea is to critically evaluate the possibility of emancipation from sufferings, following the Vedic rites and rituals.

[*]Assistant Professor, Department of Philosophy, University of North Bengal, India.

Methodology, Literature Review, Discussions and Analyses of the Subject Matter

Dichotomy of Ethical Statements

At the very outset of the discussion let us split the entire set of ethical statements into two sets, namely, the domain of public usage and that of the Vedic context. The ethical statements in the Vedas mostly speak of duties and non-duties of individuals belonging to certain sections and particular stations of life. The ethical sentences of ordinary parlance, like, 'Always speak the truth', 'Do not steal' etc. inhere a power to direct individuals accordingly, yet, they differ greatly from sentences of the Vedic realm, which speak of ritualistic duties and non-duties, in the Vedic context only. However, at times we find Vedic sentences acting as the cause of our inclination towards an action, or refraining us from an activity which is of empirical nature as well, as in '*mā gṛdhaḥ kasyasviddhanam*' (Roer 2011), meaning, we should not be jealous about others' properties. Again in other sentences like, '*na kalañjaṁ bhakṣayet*' (Bhargananda 1411 Bengali year), we find restrictions on intake of red garlic, or meat of an animal killed by a poisonous weapon. In the present context, the initial part of our discussion would precisely be about the linguistic analyses of the Vedic sentences with regard to the Pūrva Mīmāṁsā tradition, in context to their potency of moral communication, and their power to determine the ethics of various acts related to Vedic rites and sacrifices.

Vedas as the Ultimate Sanction

According to the Indian orthodox philosophical traditions, Vedic sentences are taken to be the sources of knowledge of the domain which lies beyond our sense-experience. Whatever be the content of the injunction, empirical or supra-empirical, the compulsive force expressed through the optative or potential mood of the verb (denoting sense of duty) leads us to act, or to refrain from, accordingly. The importance of the Vedic injunctions lies in the fact that they constitute such a portion of the Vedas that conveys matters not known or knowable otherwise. The injunctions enjoin such matter that is not established by any other means of proof, and enjoining thus, it becomes useful and necessary (Sukthankar 2005). To explicate further, let us consider the following injunction, – '*agnihotraṁ juhuyāt svargakāmaḥ*', meaning that, one desirous of attaining heaven must perform the *agnihotra* sacrifice. It is evident that such knowledge is never obtainable by other means, like, the perception, inference, comparison, presumption *etc*., which are strictly the means of attaining cognitions of empirical verifiability only. Hence, we must admit that the Vedas are the sole sources of knowledge on matters arguably beyond the empirical and/or sensory realm. Consequently, the Vedic injunctions are considered to be the supreme sanctions in the realm of supra-sensuous matters[1].

[1]'*tatrājñātārthajñāpako vedabhāgo vidhiḥ/ sa ca tādṛśaprayojanavadarthavidhānenarthavān yādṛśaṁ cārthaṁ pramāṇāntareṇāprāptaṁ vidhatte – yathā "agnihotraṁ juhuyāt svargakāmaḥ" iti vidhirmānāntareṇāprāptaṁ svargaprayojanavadhomaṁ vidhatte/*' (Sukthankar 2005, p. 16).

The Vedas, also referred to as *Śruti*, are the ultimate source of verbal testimony. According to the Pūrva Mīmāṁsā philosophers, the Vedas are not created. They are self-manifested, without a beginning, eternal and un-authored. The Vedas consist of millions of sentences which are passed on verbally through ages from a preceptor to his pupils. The Vedas provide us knowledge of that which is not knowable otherwise, and one such knowledge is that of *dharma*. The term '*dharma*' originating from the Sanskrit root verb '*dhṛ*', meaning to hold or to sustain, refers to ethics or morality. That is, morality is such that sustains human existence on earth, and the entire creation as such. It strictly refers to ethical codes and performances of morally sanctioned actions. Such knowledge is imparted to us primarily through the Vedic injunctions and prohibitions.

> Vedic injunctions possess undeniable persuasive power because of the fact that their authority and reliability are beyond the realms of doubt and error. For instance, injunctions like, '*yajeta svargakāmo*', meaning that, one desirous of attaining heaven must perform specific Vedic sacrifices. These dictates are neither questionable nor deceptive. The above linguistic expression does not carry any sense of probability or uncertainty in it. Śabarasvāmi, the commentator on Mīmāṁsā aphorisms, points out that the linguistic usages of human beings in the ordinary parlance, like, 'there are fruits on the bank of the river' (*Śābarabhāṣya* on *Mīmāṁsā Sūtra* 1.1.2), may be either true or false, and it is empirically verifiable. On the contrary, whatever is expressed by the Vedic sentences is not at all contradicted by the subsequent cognition of a person in a different situation, or by different individuals in different times and places. Hence, its infallibility is unquestioned and is free from all kinds of uncertainty[2].

Grammatical Analysis of a Vedic Injunction to Show its Compelling Force

Let us now try to analyse how a Vedic injunction works. In the injunction, '*yajeta svargakāmaḥ*', the verb '*yajeta*' consists of the directive power of the injunction in question. The verb '*yajeta*' is constituted of the root '*yaj*' and the suffix. The suffix again consists of two constituent parts, - the verb-notion and the optative or potential mood. We know that in Sanskrit grammar, there are ten participles signifying tense or mood[3]. The verb-notion is present in all the ten participles, but the compulsive force (signifying, sense of duty) is specific to the optative participle only. The conjugated sense of both the parts of the suffix gives rise to a creative energy towards performing actions. Before the production of an action, the favorable factor, that is, a kind of intent or mental propensity of the individual which propels the action is called the creative energy (*bhāvanā*)[4]. The creative energy is again of two types, - verbal creative energy and end-creative energy (Gajendragadkar and Karmarkar 1998). The verbal creative energy is

[2]The indubitable and infallible nature of Vedic statements are expressed as, - '*na ca svargakāmo yajeta ityatovacanāt sandigdhamavagamyate bhavati vā svargo na vā bhavatīti/na ca niścitamavagamyamānamidaṁ mithyā syāt... na ca iṣa kālāntare puruṣāntare' vasthāntaredeśāntarevā viparyeti/ tasmādavitathaḥ//*' (*Śābarabhāṣya* on *Mīmāṁsā Sūtra* 1.1.2).

[3]The central force of persuasion which remains embedded in the Vedic injunctions lies in the verb '*yajeta*' in the above example and is discussed in much detail in Sukthankar (2005, p. 11 onwards).

[4]'*ubhābhyāmapyaṁśābhyāṁ bhāvanaiva ucyate/bhāvanānāmabhaviturbhavanānukūlo bhāvayiturvyāpāra viśeṣaḥ/ sādvedhā, śābdībhāvanā, ārthībhāvanā ceti*' (Sukthankar 2005, p. 11).

responsible for producing the compulsive conducive force in the mind of the agent for the performance of the action. It is expressed by the optative element in the suffix. The end-creative energy is the actual or the final end-oriented activity which leads to the accomplishment of the action. The end-creative energy is preceded by verbal creative energy. In other words, the verbal creative energy produces the end-creative energy. We might consider an everyday experience to explicate the above notions. For instance, X is cooking rice. This action is preceded by the creative energy of X which leads to the performance of the action. Now, Y asks X to cook rice. X first listens to the instructive sentence. The words in the instruction, accordingly, produce an inspiration in the listener's mind (here, X) and it is known as verbal creative energy. At the next moment, the meaning of the words along with the end-notion produces yet another inspiration in her/his mind. Thus, the activity is about to be undertaken in order to accomplish the end. This is referred to as end-creative energy.

Analogously, in case of Vedic injunctions like, '*yajeta svargakamaḥ*', the term '*yajeta*' is responsible for the production of both verbal creative energy and end-creative energy successively in the individual's mind in a similar manner, which then leads to the production of inclination in the listener. The inclination is of the form, 'This Vedic injunction is inspiring me to perform the sacrifice'. Thus, it might be claimed that the persuasive power of the Vedic injunctions is communicated to the individual through a psychological experience, namely, creative energy, and the entire force inheres in the suffix in the form of the compulsive force signifying the sense of duty, that is, through the optative or potential mood. Due to this reason, the injunctions are considered as sentences which inspire to act. However, it is important to mention here that there is a difference between the compulsive force of injunctions and that of empirical imperatives or mere commands. The optative or potential mood of the injunctions, discussed above, does not merely act as a propeller or stimulus. Rather there is a rigorous sense of obligation entwined with it.

Vedic Prohibitions

Now let us turn our attention to another significant part of the Vedas, namely, the prohibitions. We know that just as human beings desire to attain pleasure, they also want to avoid pain and misery. This is specifically where the prohibitory statements play their active roles. These instructive statements prevent us from performing certain acts which may bring about misery and pain upon us. Thus, they are regarded as statements which refrain us from actions. These Vedic sentences, in turn, are equally powerful in communicating and strictly preventing acts like the enjoyment of prohibited objects out of sheer passion, as it would inevitably lead to extreme sufferings[5].

[5]'*puruṣasya nivartakaṁ vākyaṁ niṣedhaḥ, niṣedhavākyānāmanarthahetu kriyā nivṛttijanakatvenaivārthavatvāt/ tathāhi, yathā, vidhiḥ pravartanāṁ pratipādayan svapravartakatvanirvāhārthaṁ vidheyasya yāgāderiṣṭasādhanatvamākṣipan puruṣaṁ tatra pravartayati/ tathā* '*na kalañjaṁ bhakṣayet*' *ityadi niṣedho'pi nivartanāṁ pratipādayan*

One such instance of the Vedic prohibitions is, '*na brāhmano hantavya*' (Bhargananda 1411 Bengali year), meaning that one should not kill a Brahmin, that is, a person who might be a possessor of supreme knowledge, or a servant of God. Such prohibitory statements again carry a sense of obligation, and thus, on hearing such sentences, more often than not, individuals refrain from committing such acts.

It is, thus, evident that the Vedic injunctions inspire an individual to perform rites and rituals in accordance with one's desire and one does so out of the sense of obligation as imposed on them by optative or potential mood of the injunctions. Similarly, the Vedic prohibitions prevent one from involving in prohibited acts with the same intensity as the injunctions. The injunctions and prohibitions respectively communicate the senses of 'duty' and 'non-duty' to an individual. One realizes what *should* be done and what *should not* be done. Clearly, this is nothing but the basic tenet of a moral prescription.

Are the Vedic Means Soteriological?

The Vedas being the ultimate source of Soteriology, at this point, let us examine the efficacy of the Vedic prescriptions in order to attain emancipation from sufferings. All individuals on earth strive to attain cessation of sufferings through various means. All human endeavours are directed towards such. Thus, a prudent being would strive to attain such cessation of sufferings which would be sanguine and that which would never recur. We know that such cessation of sufferings is never attainable by empirical means, since, those means are not strong enough to prevent the recurrence of sufferings. Hence, reflective individuals would always resort to such paths which would lead them to their desired ends. That is, adopting scriptural means would provide us with the means of absolute cessation of sufferings. In Sāṁkhya philosophy, the most ancient orthodox school of thought in the Indian philosophical tradition, we find some intriguing critiques regarding the tenability of the Vedic rituals as a means of emancipation from sufferings. The Vedic means are mostly admitted by the Pūrva Mīmāṁsā philosophers or the Karmavādī Mīmāṁsā philosophers. According to the Sāṁkhya philosophers, though liberation is identical with the absolute cessation of sufferings, but it can only be attained by acquiring discriminatory knowledge (*vivekajñāna*) between the consciousness (*puruṣa*) and the matter (*prakṛti*). In case of an embodied consciousness, there appears an apparent non-apprehension of the distinctive cognition between the psycho-somatic states of the individual and the being as pure consciousness. This precisely acts as the root of all our sufferings, and hence, to liberate oneself from such induced bindings permanently, one has to attain discriminatory cognition (*vivekakhyāti*) between the consciousness (*puruṣa*) and the matter (*prakṛti*).

On that note, as discussed so far, we find that the means stated by the Vedas are similar to the empirical methods of pain eradication. That is, they cannot bring about absolute cessation of sufferings. As elaborated above, there are various

---

svanivartakatvanirvāhārthaṁ niṣedhasya kalañja bhakṣaṇasya parāniṣṭasādhanatvamākṣipan puruṣaṁ tato nivartayati' (Bhargananda 1411 Bengali year, p. 140).

sacrifices, rites and rituals mentioned in the Vedas which help to fulfil the different kinds of desires of beings, and thus, remove sufferings. These include sacrifices like *jyotiṣṭoma, agnihotra, aśvamedha, viśvajit etc.,* which respectively fulfil the purposes of attaining heaven, victory over enemies, acquiring huge areas of land and thus widening the territory of the kingdom *etc.* All these apparently are responsible for removal of sufferings, as they bring about immense pleasure as their respective consequences, as per the desire of the individuals. However, the question remains that whether such freedom from sufferings is eternal or not. That is, whether ends like heaven can provide absolute cessation of sufferings or not. According to the noted commentator Vācaspati Miśra[6], heaven involves that state of pleasure which is not only unstinted by sufferings, rather it is contradictory to sufferings (Goswami, 1406 Bengali year). Further, Kumārila Bhaṭṭa[7] states that the unstinted and the extreme pleasure which is to be enjoyed, can only be experienced in a place which is devoid of conflicts and contradictions (Goswami 1406 Bengali year). In the empirical world we can never find such a place which is free from contradictions, even momentarily. Thus, the unstinted pleasure called heaven, can only be experienced at a particular place, that is, the abode of the deities (Goswami 1406 Bengali year); though it is to be noted that traditionally heaven is accepted by most philosophers as a *state of being*, and not as a place[8].

Debate between Karmavādī Mīmāṁsā Philosophers and Their Opponents on the Nature of Liberation

There is a section of the Mīmāṁsā system called the Karmavādī Mīmāṁsā philosophers who admit heavenly bliss to be the highest end of human life, and they establish the eternal character of the heavenly bliss on the basis of certain scriptural statements[9]. Reference to the views of these philosophers may be found in the verses of the second chapter of *Śrimadbhagvadgītā*[10]. However, the mainstream schools of the Mīmāṁsā system, such as the Bhāṭṭa Mīmāṁsā school, do not admit this view of the Karmavādī Mīmāṁsā philosophers. The Sāṁkhya philosophers also do not consider heaven to be eternal. The philosophers who do not admit the eternal character of heavenly bliss establish their thesis on the basis of both inference and scriptural statements. The inference which they employ to establish their thesis is as follows – '*vimataḥ svargaḥ anityaḥ kṛtakatvāt ghaṭavat*'. It means that, heaven or heavenly pleasure is non-eternal, because it is created, just

[6]'*duḥkhavirodhīsukhaviśeṣaśca svargaḥ*', Miśra, Vācaspati in Goswami (1406 Bengali year, p. 20).
[7]'*yā prītiḥ niratiśayā, anubhavitavyā/ sā cā uṣṇaśītādidvandvarahite deśe śakyā anubhavitum/ asmin ca deśe muhūrtaśatabhāgaḥ api dvandvai na mucyate/ tasmāt niratiśaya-prītyanubhavāya kalpyaḥ viśiṣṭa deśaḥ//*', Miśra, Vācaspati in Goswami (1406 Bengali year, p. 19).
[8]'*yanna duḥkhena sambhinnaṁ na ca grastamanantaram/ abhilāṣopanītaṁ ca tatsukhaṁ svaḥpadāspadam//*', Miśra, Vācaspati in Goswami (1406 Bengali year, p. 19).
[9]'*apāma somamṛtā abhūma*' in Goswami (1406 Bengali year, p. 20).
[10]'*yāmimāṁ puṣpitāṁ vācaṁ pravadantyavipaścitaḥ/ vedavādaratāḥ pārthanānyadastīti vādinaḥ//*', *Śrīmadbhagavadgītā* Verse 42 (Brahma 1986).
'*kāmātmānaḥ svargaparājanmakarmaphalapradām/ kriyāviśeṣavahulāṁ bhogaiśvaryagatiṁ prati//*', *Śrīmadbhagavadgītā* Verse 43 (Brahma 1986).
'*bhogaiśvaryaprasaktānāṁ tayāpahṛtacetasām/ vyavasāyātmikābuddhiḥ samādhaunavidhīyate//*', *Śrīmadbhagavadgītā* Verse 44 (Brahma 1986).

like a jar. We know that if something is created or has a beginning, it would have an end too. Thus, things with a beginning, presuppose an end, and hence, are non-eternal. The state of heavenly pleasure is produced as a result of the sacrifices, and thus, it is generated, just as a jar. Accordingly, it cannot be eternal.

The Karmavādī Mīmāṁsā philosophers might argue that this inference is not capable of establishing the non-eternal character of heavenly pleasure. The eternal character of heavenly pleasure is established by the scriptures themselves. It is said that the deities performed the Vedic rite called *soma yāga* and drank the extracts of the creeper called *soma*, and as a result, they attained the status of 'deathlessness' (*amṛtatva*). Now the term '*amṛta*' etymologically means deathlessness or going beyond death. Thus, the deities went beyond death or transcended death by performing this rite and became eternal. Since the deities are eternal, the heavenly bliss or heavenly pleasure enjoyed by them is also eternal. This scriptural statement overrides the above inference by which other philosophers have tried to establish the non-eternal character of heavenly pleasure.

Against the above defence of the Karmavādī Mīmāṁsā philosophers, the Sāṁkhya philosophers point out that whenever there is a conflict or contradiction between a scriptural statement and some other valid epistemic instrument, it cannot always be said that the scriptural statement is stronger than the other instruments. This is due to the fact that a perceptual cognition and an inference cannot establish its object in any way other than it actually does. If the scriptural statement is always considered as stronger than the other epistemic instruments, then one would have to forego or give up some other valid epistemic instrument, such as a veridical perception or a valid inference. However, if the validity of valid epistemic instruments is denied, then the entire epistemology will lose its trustworthiness. For this reason, no orthodox Indian philosopher places any kind of blind trust or credence upon a scriptural statement, whenever there is any conflict between the scriptures and other epistemic instruments. Rather the classical Indian philosophers evaluate the relative strength and weakness of each epistemic instrument and only such assessment of relative strength can determine whether a particular epistemic instrument can override another. On that note, the Sāṁkhya philosophers talk about a specific methodology to resolve the above conflict between the eternal and non-eternal character of heaven, as endorsed by the scriptures and denied by the opponents, respectively. This is being discussed in the following section.

Sāṁkhya Critique of the Karmavādī Mīmāṁsā Philosophers' View

i) Methodology and arguments from the paradigm of Sāṁkhya philosophy

To prove the non-eternal character of heavenly pleasure, the Sāṁkhya philosophers employ the methodology called argumentation on the statements of wider and narrower scope (*sāvakāśa-niravakāśa nyāya*). The Mīmāṁsā system invented the said method of argumentation to resolve conflicting scriptural statements. It states that whenever there is a conflict between two rules or two

scriptural statements of which one has a wider scope than the other, the rule or the statement of lesser scope should be considered as stronger than the rule or the statement of the wider scope. For, if the rule or the statement having the wider scope is considered as stronger then, in every instance, it will be so, and consequently, the statement or the rule having the smaller scope would have no scope at all. In that case the rule or the scriptural statement having the smaller scope would not have any application at all, and hence, could not be regarded as a veracious statement or rule at all, even in its specific context. The matter is clarified by the Sāṁkhya philosophers by referring to another employment of this methodology. For instance, the scriptural statements – '*mā hiṁsyāt sarvābhūtāni*' (meaning, never harm any being) and '*agniṣomīyaṁ paśumālabheta*' (meaning, sacrifice animals for the performances of *agni* and *soma* sacrifices) contradict one another. This is because the former statement forbids violence towards any organism, whereas the second prescribes animal sacrifice for appeasing the deities, like *agni*, in case of sacrifices. Here the former statement obviously has a wider scope than the latter. Now if the statement with the wider scope, that is, the first statement is considered as stronger than the latter statement, then the latter would not be applied at all. Hence, it cannot be treated as a source of veridical cognition. Now if the validity of one Vedic statement is denied, then the entire Vedas would be at the risk of losing its supreme authenticity. For this reason, in this context, the statement with the smaller scope, that is, the latter statement is considered as stronger than the statement of the wider scope. In that case, the latter statement would mean what it literally means, but the significance of the former statement would have to be restricted in conformity with the significance of the latter statement. Thus, the former statement would mean that violence is forbidden in all other cases, except in case of certain sacrifices. When thus interpreted, neither of the two statements loses their validity and the veracity of the entire Vedas also is not subjected to doubt. Similarly, whenever a valid inference is at odds with any scriptural statement, the inference is considered stronger because of its smaller scope than the scriptural statement. This is because, in an inference the probandum is established in the locus of the inference with the help of a probans, and it cannot be established otherwise. On the contrary, a scriptural statement being a linguistic entity can be interpreted in many different ways. So the scope of a scriptural statement is always greater than a veridical perception or a valid inference. For this reason, whenever a veridical perception or a valid inference comes into conflict with a scriptural statement, the scriptural statement is interpreted in accordance with the perception or the inference. The inference, as stated above, which demonstrates the non-eternal character of heavenly pleasure is a valid one. So the validity of this inference cannot be denied. Hence, the term '*amṛta*' occurring in the statement '*apāma somamṛtā abhūma*' must carry some other sense. In fact, this other meaning is indicated by another scriptural statement in the *Viṣṇupurāṇa*, - '*ābhūtasaṁplavasthānamamṛtatvaṁ hi gīyate*', meaning that the heavenly bliss lasts till the destruction of a particular creation. So compared to ordinary pleasures, heavenly bliss or heavenly pleasure lasts for a long time, but it is not everlasting or eternal. The Vedic means are vitiated by the defect of erosion or destructibility. It is to be noted here that Īśvarakṛṣṇa, the author of *Sāṁkhyakārikā*, and other

Sāṁkhya philosophers are not talking about the destructibility of the means. This is because a Vedic rite being an action is obviously of a particular duration. So Īśvarakṛṣṇa here is talking about the destructibility or the non-eternal character of the end which is attained through these Vedic rites and this end is nothing but heavenly pleasure. Since, the end attained through the Vedic means may be destroyed; the Vedic means are at par with the empirical means of overcoming sufferings, such as medicines *etc.*

ii) Discussion on the nature of heaven

Now there remains an apprehension that due to the presence of the causes of sufferings, one might be afflicted by pain even after the attainment of heaven. In that case, the attainment of heaven might not be the desired end of the individual. To resolve such discomfort, Vācaspati Miśra says that heaven is that kind of pleasure which is not only contradictory to sufferings; rather it destroys all kinds of pain which are impediment to it. Further, it also eradicates all the causes of sufferings, including the root cause, that is, consequences of one's actions from previous lives[11]. He further claims that heaven is not something which would erode after a point of time[12]. However, the opponents might argue that heaven being the result of Vedic sacrifices, is a positive entity which is produced, and hence, it cannot be eternal[13]. In refutation of the above, the Vedic tenet which has been cited is – '*apāma somamamṛtā abhūma*', which means that one who drinks the *soma*, that is, the person performing the Vedic sacrifice attains deathlessness, meaning that the individual transcends death. This further establishes that whatever is produced as a result of the sacrifice, that is indestructible. Thus, from the above standpoint the advocates of the Vedic means of pain eradication argue that the method of attaining discriminatory cognition between the matter and consciousness, as held by the Sāṁkhya philosophers, is extremely difficult to achieve, as it requires the effort and care on the part of the individual over multiple lives. In contrast to that the Vedic means are easier and involve much less effort as regards the performance of the sacrifices. Thus, one should adopt the Vedic means of rites and rituals in order to remove sufferings.

iii) Vedic means are comparable to empirical means in terms of removal of sufferings

In response to the entire above discussion and the objection raised thereafter, Īśvarakṛṣṇa states in the second *Sāṁkhyakārikā*[14] that the Vedic means are similar to the empirical means in terms of the fact that the Vedic sacrifices too are not

[11]'*sa ca svasattayāsamūlaghātamapahanti duḥkham*', Miśra, Vācaspati in Goswami (1406 Bengali year, p. 20).
[12]'*na ca eṣa kṣayī*', Miśra, Vācaspati in Goswami (1406 Bengali year, p. 20).
[13]'*svargaḥ kṣayī utpattimat bhāvapadārthatvāt aihikasukhavat ghaṭapaṭādivat*', Miśra, Vācaspati in Goswami (1406 Bengali year, p. 21).
[14]'*dṛṣṭavadānuśravikaḥ sa aviśuddhikṣayātiśayayuktaḥ/ tadviparītaḥ śreyān vyaktāvyaktajña vijñānāt//*', *Sāṁkhyakārikā* 2.

capable of providing absolute emancipation from sufferings. The Vedic means of eradication of sufferings refer to the kind of knowledge which can be known after listening to the Vedic tenets from the preceptor. However, the knowledge attained thereby, that is, the cognition of the Vedic sacrifices, is analogous to the empirical means of pain removal as they are neither the exclusive means, nor can they sanction the non-recurrence of sufferings. Now one might argue that the notion of the discriminatory knowledge (*vivekajñāna*) between the consciousness (*puruṣa*) and the matter (*prakṛti*) is also obtained from the Vedas. Hence, similar to the other means which are Vedic in nature, *vivekajñāna* too cannot ensure the absolute cessation of sufferings. In response to the above apprehension, it is stated that in spite of the fact that *vivekajñāna* - the highest form of knowledge as held by the Sāṁkhya and the Yoga schools of philosophy, though endorsed by the Vedas, are not tainted with limitations, unlike the Vedic sacrifices. The following inference helps to show the limitations of the Vedic sacrificial means – '*vaidikaḥ upāyaḥ dṛṣṭatulyaḥ na aikāntikātyantikaduḥkhatrayapratikāropāyaḥ aviśuddhiyuktatvāt kṣayayuktatvāt atiśayayuktatvāt ca*' (Goswami 1406 Bengali year). That is, the Vedic means are merely at par with the empirical means of alleviating sufferings, because the Vedic sacrifices and their results are vitiated by three defects, namely, impurity, erosion and differences in magnitude. That is, the Vedic rituals often involving animal sacrifices are tainted with the quality of impurity. Also, the results that are produced as consequences of the sacrifices, however, pleasurable they might be, have an end. And, the results obtained because of the sacrifices are not of the same quality. They vary in magnitude, duration and intensity. Hence, they lack the marks of eternal bliss, or absolute cessation of sufferings.

In the above inference the locus is the Vedic means, and the probandum is the comparability of the sacrificial means with the empirical means of eradication of sufferings. However, simply stating 'Vedic means' includes the sacrifices as well as the discriminatory cognition in its purview. Thus, the afore-mentioned apprehension gets revoked. To clarify, Vācaspati Miśra states that the reference to Vedic means in the aphorism refers to the sacrificial rites and rituals only. Though it is true that the discriminatory cognition between the self or the consciousness and the matter is known from the Vedas too, yet there are Vedic statements, which according to Vācaspati Miśra, should be interpreted as the self or the consciousness is to be perceived as radically different from matter[15]. When such distinctive perception is produced, it is to be practiced over ages, and over multiple lives of an individual. Then finally sufferings are eradicated in such manner that they can never recur (*Chāndogya* 8.15)[16].

iv) **The nature of distinctive cognition: Sāṁkhya claim**

As mentioned earlier, that the results of Vedic sacrifices cannot be eternal due to the presence of impurity, erosion and degrees of variation, but such is never the case with distinctive cognition. It is devoid of and is essentially radically different

[15]'*ātmāvā'are draṣṭavyaḥ*', and '*prakṛtitaḥ vivekatavyaḥ*', Miśra, Vācaspati in Goswami (1406 Bengali year, p. 23).
[16]'*na sa punarāvartate*'.

from the above three features as we find in case of empirical pleasures and that of the form of heaven. Hence, the Sāṁkhya philosophers claim that the discriminatory cognition between the self or consciousness and the matter is more fundamental and most efficacious in terms of eradication of sufferings. Here, a few points need elucidation regarding the Sāṁkhya and the Yoga philosophy on the nature of liberation of the self. In Sāṁkhya philosophy, the internal sense organ consists of three components – namely, the intellect, the ego and the mind. The three parts have three fundamental functions in attaining knowledge of objects in the empirical realm, like a jar, a piece of cloth *etc*. and that of the self as in case of I-cognitions, like, 'I am happy', 'I am the knower' *etc*. In Sāṁkhya and the Yoga philosophy it is admitted that the intellect, the ego and the mind, along with the body are products of evolution of matter, whereas the consciousness does not evolve. It is ever pure, immutable, eternal and essentially unrelated to anything around. However, in case of beings, that is, while looking at embodied consciousness, we find that there is an apparent non-apprehension of the distinction between the self and the products of matter, namely, the intellect, the ego, the mind and the body. The Sāṁkhya epistemology holds that the sense organs like, the eyes *etc*. receive the form of the object and sends it to the internal sense organ. The intellect-ego-mind compound then takes up the form of the object, say, a jar, and through subsequent steps the determinate cognition of the form, - 'I know that this is a jar' is produced. Similarly, for instances of emotive and other cognitive states, expressions like, - 'I am happy', 'I am sad', 'I am healthy', 'I know', 'I perceive', 'I enjoy' *etc*. occur. However, all these states are modifications of the internal sense organ compound only, and none of it actually touches the consciousness. The self or the consciousness remains as it is in its pure, unaltered state. The issue that arises at this point is that the consciousness being in proximity to the internal sense organ gets reflected on the intellect which is of luminous nature and possesses the property of reflection like a mirror. This is due to the predominance of the buoyant and luminous perpetual quality (*sattva guṇa*) in it. The reflected consciousness then appears to identify itself with the modifications of the internal sense organ. Accordingly, the emotions and the cognitions that are actually there in the internal sense organ, appears to be there in the consciousness. The actual consciousness, though, remains as it is, unaffected by any of the above. This can be explained by a metaphor, as follows – if we look at a mirror which has got some spots on it, the face that appears on the mirror also carries those spots. The actual face though, remains as it is. It is only the reflected face that carries the spots from the mirror. Such is the case of the apparent non-distinction between the consciousness and the matter. It is, thus, the very objective of any prudent individual to free oneself from apparent notions, and to realise things as they are. Hence, the ultimate cognition is case of Sāṁkhya and the Yoga philosophy is to realise the consciousness as it is, and to be freed from any impact of matter whatsoever. Thus, the highest form of liberation for the Sāṁkhya and the Yoga philosophical schools is the attainment of distinctive cognition, such that the consciousness exists only in its true essence. As a result, without even any apparent relation with matter, the question of sufferings does not arise at all.

Conclusions

On arriving at the terminal part of our discussion, it may be stated that the above debate between the different schools of philosophy originates from the fact that there is non-concurrence among the views of the schools regarding the nature of liberation. Though each of the schools consider the eradication of all kinds of sufferings from the empirical realm, yet there remains disagreements regarding the nature of the ultimate existence of the self, and hence, the form of emancipation. In the noted Sanskrit text, *Sāṁkhyatattvavivecana* of Kṣemendra[17], we find that the Sāṁkhya philosophers are advocates of such liberation which is a state devoid of any attributes (*abhāvamokṣavādin*s). That is, liberation for them is the state of consciousness-as-it-is, and hence, not qualified by any other state of feelings or being, like pleasure *etc.* Such kind of absolute cessation of sufferings is considered as liberation by the Sāṁkhya philosophers, where the consciousness remains in isolation, oblivious of anything around. This is the highest and the most sought after state of the consciousness, which is technically referred to as *kaivalya*, meaning isolation. However, philosophical schools like the Pūrva Mīmāṁsā philosophers or the Karmavādī Mīmāṁsā philosophers consider liberation to be not only the cessation of sufferings, but the state of realisation of eternal pleasure. For them, the state of liberation is marked by the absence of all kinds of sufferings and is qualified by eternal bliss (*sukhamokṣavāda*). Thus, according to them the attainment of heaven itself is the eternal and unwavering state of bliss. Sāṁkhya philosophers, however, point out that the realisation of pleasure, whether it is eternal or fleeting, the realisation itself being a cognitive state is always non-eternal. Hence, liberation is not a state of realisation of eternal pleasure; rather it is the state of absolute cessation of all kinds of sufferings (Dvivedi 1920).

Now keeping the contention of our present paper in mind, we can assert that since, the Pūrva Mīmāṁsā philosophers or the Karmavādī Mīmāṁsā philosophers consider liberation to be not only the cessation of sufferings, but the state of realisation of eternal pleasure or bliss, then according to the school of thought in question, the Vedic sacrifices, and accordingly, the duties and non-duties are the sole means of attaining liberation. Accordingly, the Vedic scriptures pose to be the absolute sanction of Soteriology in the arena of such philosophical thoughts. However, it is to be kept in mind here that the Vedic means might often bring about tremendous sufferings too due to the sin incurred by performing certain sacrifices like, *śyena*, which can accomplish killing of one's enemies through sacrifices. Moreover, the performances of the sacrifices themselves are, in some way or the other, afflicted with some amount of sufferings, in spite of the fact that

[17]'*darśanaśaktirahitasya kriyāśaktimataḥ pradhānasyāpi puruṣeṇa saṁyogaḥ mokṣārtha puruṣasya bhinnatvena vyaktāvyaktapuruṣajñāne jāte pradhānasya mokṣo bhavati/ nityasukhopalabdhirmokṣa iti cedupalabdherapi nityānitya vivekagrastatvādasāram/ na ca nityasukhagocarasyāvidyādi yatkiñcidāvaraṇabhaṅga eva puruṣārthe/ vācyaḥ sukhānubhavasyaiva puruṣārthatvāccaitanyanityatvenāvaraṇasyāpi asambhavācca/ mokṣe paramānandaśrutismṛtayastu - mokṣaśāstraparibhāṣāmātrā/ duḥkhamevāsti na sukhaṁ yasmāt tadupalabhyate/ duḥkhārtasya pratīkāre sukhaṁ sajñāvidhīyate// duḥkhaṁ kāmasukhāpekṣā sukhaṁ duḥkhātyayaḥ smṛtaḥ/ ityādismṛtibhirduḥkhanivṛttireva sukhatvena paribhāṣitā/*', Kṣemendra, in Dvivedi (1920 p. 33).

they produce immense pleasure by fulfilling the desired ends. However, it is never the case with the advocates of the opposing theories, as discussed above. Thus, the Sāṁkhya philosophers argue that the Vedic rites and rituals are merely means of eradication of sufferings, if at all, in the empirical realm, but not the means of attaining liberation of the spirit in the absolute sense.

Finally, it may also be highlighted that the Vedas though prescriptive in nature, are of the structure of hypothetical imperatives. This is due to the fact that the injunctions are end-specific. They posit that if one desires a particular end, then they must perform a specific sacrifice. For instance, *if* individuals desire to attain heaven, *then* they must perform a certain sacrifice. The evident *if-then* form presents the injunction as a hypothetical imperative. This portrays the fact the Vedic prescriptions, though of the nature of commands, accommodates the free will of individuals. Hence, the good or bad of actions are not merely directed by the Vedas, rather it depends upon the choices that the individuals make for themselves. Thus, it could be claimed that the Vedic ethics involves paradigms to ascertain a good life for beings, based on individual choices. At the same time, based on such action-consequence principle discussed throughout, we may further claim that it provides the means of eradication of sufferings in order to attain eternal bliss, or in other words, liberation as admitted by certain sections of philosophical schools.

References

Bhargananda S (Ed.) (1411 Bengali year) *Laugākṣibhāskaraviracitaḥ Arthasaṁgrahaḥ*. (The Compendium of Meanings composed by Laugākṣibhāskara). Kolkata: Sanskrit Pustak Bhandar.

Brahma N (Ed.) (1986) *Śrīmadbhagavadgītā with the commentary (gūḍhārthadīpikā) of Madhusūdana*. Kolkata: Navabharata Publishers.

Dvivedi V (Ed.) (1920) *Sāṁkhyasaṁgraha*. (Compendium of Sankhya). Varanasi: Chaukhamba Sanskrit Series Office.

Gajendragadkar AB, Karmarkar RD (Eds.) (1998) *The Arthasaṁgraha of Laugākşi Bhāskara*. Delhi: Motilal Banarsidass Publishers Private Limited.

Goswami NC (Ed.) (1406 Bengali year) *Vācaspatimiśraviracitā Sāṁkhyatattvakaumudī*. (The Sāṁkhyatattvakaumudī composed by Vācaspatimiśra). Kolkata: Sanskrit Pustak Bhandar.

Roer E (Trans.) (2011) *Īśopaniṣad*. In *The Twelve Principal Upanishads*. Delhi: Divine Books.

Śābarabhāṣya. (The Commentary of Śabara). In *Mīmāṁsā Sūtra* 1.1.2.

Sukthankar SS (Ed.) (2005) *Arthasaṁgrahaḥ of Shri Laugākşhi Bhāskara*. Delhi: Vista International Publishing House.

Athens Journal of Philosophy - Volume 3, Issue 3, September 2024 – Pages 149-162

The Kantian Notion of Categories and their Origin

By Dipanwita Chakrabarti[*]

The objective of the present work is to understand and elucidate Kant's notion of category and how he derived the categories from a single transcendental principle. Kant did not put forward any definition of categories. He believed that categories cannot be defined without perpetrating a circle. Thus, he began his discourse with certain features of categories in his work Critique of Pure Reason. We have discussed the characteristic features of Kantian categories. An important point to be noted here is that the categories, in the fullest Kantian sense of the term, must have a distinct property, namely that it should necessarily be applicable to all objects of knowledge. However, we are not concerned with the necessary applicability of concept to all objects of knowledge here in this paper. The analysis about the Kantian notion of categories, more importantly, necessitates a discussion about how he derived them from a single transcendental principle. Kant referred to the single principle which guides the search for the categories as "the clue to the discovery of the categories." The specific and clear formulation of the principle which served as the transcendental clue to the discovery of the categories for Kant is that to every form of judgment there corresponds a pure and basic concept of the understanding. The forms of judgments and the categories both originate from the same source, namely, the function of the understanding, i.e., thinking. It may be noted here that the understanding is the power or faculty of knowing and thinking or judging is the function of understanding. Kant argued that the twelve logical forms of judgments provided the clue to the origin of twelve corresponding a priori concepts or categories. Two arguments provided by Kant in support of the principle serving as a transcendental clue to the discovery of the categories are analysed. An orthodox view held by some philosophers that for Kant the forms of judgment are forms of analytic judgment has been critically analyzed and is interpreted as erroneous.

Keywords: *category, judgment, understanding, categorematic, syncategorematic*

Introduction

The term 'categories' in its philosophical usages is not introduced by Kant for the first time. His notion of categories, however, is markedly original. It is necessary therefore to try at the very outset to be clear about what he meant by categories. The notion of categories would not be clear unless we discuss the origin of those categories. The primary intent of the present paper is to clarify Kant's notion of categories and their origin.

Kant did not put forward any definition of categories. He believed that categories cannot be defined without perpetrating a circle. Therefore, he began his discourse with certain features of categories, instead of a definition, in his work

[*]Associate Professor, Department of Philosophy, Vidyasagar College (affiliated to University of Calcutta), India.

Critique of Pure Reason, henceforth referred to as *Critique*. An important point to be noted here is that categories, in the fullest Kantian sense, must have a distinct property, namely that it should necessarily be applicable to all objects of knowledge. However, here in this paper, we are not concerned with the necessary applicability of concept to all objects of knowledge. This aspect is discussed by Kant under the section *Transcendental deduction* and *Analytic of Principles* in his work *Critique*. The analysis about Kantian notion of categories necessitates the discussion about how and whereof did Kant derive the categories. Indeed, Kant attempted to derive all the categories from a single principle. This principle which acts as a guide to the search for the categories is transcendental in nature and is called the clue to the discovery of the categories. Kant has provided two arguments in support of the principle serving as a transcendental clue to the discovery of the categories. An attempt has also been made in the present work to provide an understanding of the objection raised against the principle which acts as a transcendental clue to the discovery of categories and study the validity of this objection in the light of Kant's own work.

The text has been divided into several sections. Each of these sections deal with the issues underlined above.

Kant on the Definition of the Categories

Kant has not provided us with any definition of the categories. He has made two different kinds of assertions on the question of such definitions.

In the course of the metaphysical deduction, Kant observed: "In this treatise, I purposely omit the definitions of the categories, although I may be in possession of them… In a system of pure reason, definitions of the categories … merely divert attention from the main object of enquiry, arousing doubts and objections which, without detriment to what is essential to our purposes, can very well be reserved for another occasion" (Kant 1978, $A_{82\text{-}83}/B_{108\text{-}109}$). This observation suggests that the categories are definable and that Kant has been possibly in possession of their definitions. But this suggestion is not in keeping with his view expressed elsewhere in the *Critique*. There he said that the categories are indefinable. Thus, he observed: (A_{245}) "But they (the categories) cannot themselves be defined. The logical functions of judgments in general, unity and plurality, assertion and denial, subject and predicate, cannot be defined without perpetrating a circle, since the definition must itself be a judgment, and so must already contain these functions" (Kant 1978, $A_{244\text{-}245}$).

It appears that according to Kant categories cannot be defined without perpetrating a circle. A definition can be given only in the form of a judgment and all judgments must involve categories. Thus, in order to define a category, we require categories. This is a viciously circular reasoning. Hence, instead of a definition Kant began with a kind of characterisation.

This later observation seems to reflect a mature view and so, we may attach a greater importance to it. The impression should not persist that the categories, which are logically indefinable, are on this account devoid of meaning. They are

as clearly intelligible as the logical forms of judgments themselves. They come to acquire a determinate meaning and relation to any object by virtue of the general condition of sensibility, that is, in so far as they are schematised. Since at present we are concerned with the pure categories, and not with the schematised ones, the question of the determinate meaning and significance conferred on the categories by schematisation need not bother us.

An objection may be raised against Kant's view that categories cannot be defined without perpetrating a circle. It is true that any definition needs a judgment and hence the use of categories. But to conclude that this would lead to a vicious form of circularity is unwarranted and itself fallacious. Using a premise in demonstration of what that premise says is circular in a vicious manner, but employing judgment and thus categories to define (not demonstrate) what categories are in general, is logically unproblematic. The function of definition is not to demonstrate anything but to add conceptual clarity to a previously obscure notion.

Kant's Characterisation of Categories

As is well known, both intuitions and concepts are, for Kant, necessary for knowledge in the sense of knowledge of objects. Through intuition an object is first given to us, through concepts the object is thought in relation to the given representation. Concepts, however, are of no use unless they are referred to the object, or rather, to the sense-manifold, presented in intuition. In other words, concepts must be used in judgments. Judgments are formed by combining or connecting certain concepts with one another. The mere combination of one concept with another does not result in a judgment. One concept may be combined with another so as to obtain a compound concept. There are certain definite rules, some basic ways in which concepts are united in judgments. Reflection on these basic ways or rules gives us certain concepts, which are necessary to any judgments, in that they are concepts required for uniting or connecting any given concepts into judgments. Such connective concepts - concepts of connexion - are categories for Kant (Dryer 1966, p.112).

Categories, for Kant, are, then, certain syncategorematic concepts which represent the basic rules or ways in which categorematic concepts are connected or united into judgments. Lewis White Beck has made the point adequately clear:

> A judgment is a synthesis of concepts according to a rule. For instance, if I should have, as it were floating around in my consciousness, the categorematic concepts 'black' and 'pipe', the rules of judgment limit the possible ways I could combine them into a unitary judgment: 'the pipe is not black', 'the pipe must be black', 'the pipe may be black', 'if the pipe is black …', but not 'the is pipe black'. Now each of these rules corresponds to a concept of a kind or form of synthetic unity in which categorematic concepts are content. To each rule there corresponds a pure concept of the understanding- pure because syncategorematic concepts such as 'all', 'some', 'not', 'if, …then', 'either… or', 'is', 'may be', and 'must be' are not derived a posteriori: from experience. They are concepts supplied by the understanding itself,

> the faculty of synthesizing categorematic concepts into judgments (Beck 1969, p. 447).

We are now in a position to appreciate Kant's own statement that categories are "original pure concepts of synthesis that the understanding contains within itself *a priori"* (Kant 1978, A_{80}/B_{106}). An examination of this statement shows that categories, according to Kant, are concepts having certain features. These are (i) concepts of synthesis; (ii) concepts of the understanding, i.e., intellectual, not sensible, concepts; (iii) they are original, i.e., basic or primary concepts, and (iv) they are pure or a priori concepts. These features of categories need to be explained.

In the first place, categories are concepts of synthesis. Part of what is meant by this has already been pointed out. Categories are concepts of synthesis, because they represent, or correspond to, the basic ways in which certain (categorematic) concepts are united or synthesised into judgments. But something more is meant than merely this when Kant says that categories are concepts of synthesis. Consideration of this additional significance may, however, be postponed for the time being.

Now, in the second place, categories are concepts of the understanding. That is, they are intellectual, not sensible concepts. This might be objected to since, for Kant, all concepts rest on the understanding and only intuitions are due to sensibility, it is a redundancy to describe the categories alone as 'concepts of the understanding' and even misleading to describe them so, because this may suggest as if other kinds of concepts were due to sensibility (Kant1978, A_{51}/B_{75}, A_{68}/B_{93}). In answer to this objection, it might be said that categories are concepts of the understanding in a special sense. All concepts are made by the understanding as regards their form only, but not as regards their matter. The matter (i.e., the content) of such concepts that are derived by the understanding from intuitions, presented by sensibility, is obviously not made by the understanding. Such concepts are, because of their matter, not concepts of the understanding but sensible concepts. Categories are to be distinguished from such sensible concepts. Both the matter and the form of the categories are due to the understanding. That is why, Kant said that the understanding contains them 'within itself'. They are related to our understanding in a special way in that they are the concepts or thoughts of which our understanding must avail itself in uniting concepts of other kinds into judgments.

In the third place, the categories are original concepts. By 'original,' Kant here meant 'underivative' (Kant 1978, A_{82}/B_{108}). Concepts derived from empirical intuitions are empirical concepts, such as those of redness, blueness, etc. Concepts derived from pure intuitions are called 'modes of pure sensibility,' e.g., the concepts of spatiality, temporality, triangularity, etc. Again, concepts derived from the combination of one category with another or with a mode of pure sensibility were called (by Kant) 'predicables.' From such derivative concepts, i.e., empirical concepts, modes of pure sensibility, and predicables, Kant distinguished his categories by dubbing them 'original.'

Doubtless, the categories stand distinguished from empirical concepts and modes of pure sensibility by the fact that they are 'concepts of the understanding,' but they cannot, merely by virtue of this fact, be distinguished from such

predicables as resulting from the combination of one category with another. Hence the qualification 'original,' added by Kant, is not superfluous.

In the fourth place, the categories are pure or a priori concepts. Some might object that this qualification is superfluous, because a priority of the categories is already indicated by the fact that they are original or underivative concepts. Empirical concepts are derivative, being derived from empirical intuitions. Hence categories, as underivative concepts, cannot be empirical and so must be a priori. This objection, however, is based on a neglect of the philosophical tradition. Empiricists like Locke and Hume regarded certain concepts as basic or original. They called them simple concepts and maintained that complex concepts are formed by compounding the simple concepts (Locke1975, Book II, Chap. II, Sec. 1; Book II, Chap. XII, Sec. 1; Hume1982, Sec. II). For them, these simple concepts are empirical and not a priori. The objectors might here retort that even the simple concepts of the empiricists are on the Kantian view, derivative, not original concepts, as being derived from sensations. This is no doubt true, but to insist on this point would be to legislate for other philosophers as to what concepts, or what kind of concepts, are to be reckoned as basic. The empiricists have indeed the right to decide for themselves that their simple concepts are basic concepts. The deeper reason remains to be considered why Kant regarded his basic concepts as a priori. Simple empirical concepts of the empiricists are not syncategorematic, but categorematic. They can be, and often are, used as terms in judgments. Accordingly, they also are subject to the Kantian categories, which are concepts required for uniting any categorematic concepts into judgments. Categories are a priori, because they come to be known only through reflection on the ways in which any given concepts are used in judgments, and not by appealing to empirical observations.

We may now redeem a promise made earlier. We said that we should say something more on the point that the Kantian categories are 'concepts of synthesis.' Categories are concepts of synthesis, not merely because (as already mentioned) they represent the ways in which any given concept is connected in judgments, but also because, (as is to be emphasised now) they represent the ways in which the given indeterminate manifolds of sense are necessarily organised or synthesised into determinate objects of knowledge. Emphasis on this additional point does involve a tacit reference to Kant's Copernican Revolution in philosophy. The term 'category,' as used by Baumgarten before Kant, stands for the universal predicates (Paton 1936, p. 257). If Kant has chosen to apply this term to the basic pure concepts of the understanding, it is because he thinks that such concepts apply universally and necessarily to objects of thought and even of knowledge. The universal and necessary applicability of categories to objects of knowledge is for Kant guaranteed by their being concepts of synthesis. The categories represent certain conceptual rules in accordance with which the sense-manifold must be, on Kant's view, synthesised into determinate objects of knowledge, and accordingly the categorial features must universally characterise the objects of knowledge. Now, in order to make the notion of categories clear it is necessary to know how and wherefrom Kant derived the categories or the pure concepts of the understanding. What is the clue to the discovery of the categories?

Kant's Attempt to Trace the Origin of Categories

Kant put forward his claim in his metaphysical deduction of the categories in the *Critique*. He said:

> [The] following are the points of chief concern: (1) that the concepts be pure and not empirical; (2) that they belong, not to intuition and sensibility, but to thought and understanding; (3) that they be fundamental and be carefully distinguished from those which are derivative or composite; (4) that our table of concepts be complete, covering the whole field of the pure understanding (Kant 1978, B_{89}).

Kant's objective in the metaphysical deduction was not to prove that certain concepts were categories in the fullest sense that the term category bears in the Kantian philosophy. Kant's own claim is comparatively modest.

It is clear that Kant, in his metaphysical deduction, was concerned with showing that certain concepts were pure, intellectual, and fundamental, and that they together constituted a complete system of the basic elements of the understanding. If any concept can indeed be shown to be such, it would be fair to maintain that they lay a good claim to the categorical status.

However, it does not follow that concepts that possess the above-mentioned features, are necessarily categories in the fullest Kantian sense. Categories, in the fullest sense, must have an additional property, namely, necessary applicability to all objects of knowledge.

The problem of necessary applicability of certain concepts to all objects of knowledge is tackled by Kant in his *Transcendental Deduction and Analytic of Principles*. This topic has not been discussed in the present work.

We may, therefore, say that Kant's argument to prove that certain concepts are categories in the fullest sense was a complex argument that began from the metaphysical deduction and developed progressively through the transcendental deduction till the analytic of principles. Kant's task in the metaphysical deduction was not simply to discover the categories in the sense of making known certain concepts which were previously unknown. The task was rather to trace the origin of the categories to their common source in the nature of the understanding according to a principle. It is this task which cost Kant several years of hard reflection.

There was a time when Kant, like many others, put the ideas of space and time on the same footing with such other concepts like those of existence, possibility, necessity, ground, unity, plurality, etc... It was after a long reflection that he came to distinguish the categories as basic concepts of the understanding from the elementary notions of sensibility such as space and time. Thus, he wrote in the *Prolegomena*:

> After long reflection on the pure elements of human knowledge, (those which contain nothing empirical), I at last succeeded in distinguishing with certainty and in separating the pure elementary notions of the sensibility (space and time) from those of the understanding (Kant 1950, p. 70).

The distinction between intellectual concepts and sensible ideas came to be made by him for the first time in 1770 in his *Inaugural Dissertation.* We find him observing there: "(The) concepts met with in metaphysics are not to be sought in the senses, but in the very nature of the pure intellect, … To this genus belong possibility, existence, cause, etc., together with their opposites or correlates. These never enter any sensual representations as parts…" (Kant 1967, p. 59). As for the ideas of space and time, he said that these were not concepts but pure intuitions in this work and that these originate in our sensibility, not in the understanding. This view was never changed by him subsequently, and was in fact reasserted in the section *Transcendental Aesthetic* of the *Critique.*

The distinction of the categories from the basic notion (intuitions) of sensibility being fixed, Kant's next task was to derive the categories from the understanding according to some principle or principles. In his famous letter of 21st February 1772 to Marcus Harz, he expressed the hope that the categories could be derived from the understanding in accordance to a few principles or, as he said, by "following a few fundamental laws of the understanding (Kant 1967, p. 73)." He did not explain, however, what these few fundamental laws or principles were.

But Kant could not remain satisfied with the position reached in 1772. In the *Critique*, he abandoned the idea that categories were to be derived according to a few principles. Here, he insisted on the need for a single principle so that the categories might be reduced to a system (Kant1978, A_{67}/B_{92}) and not merely to classes. As he said in the *Prolegomena*. "There can be nothing more desirable to a philosopher than to be able to derive the scattered multiplicity of the concepts or principles which had occurred to him in concrete use from a principle a priori, and to unite everything in this way in one cognition…. This constitutes comprehension; and only then has he attained a system (Kant 1950, p. 69 f)."

According to Kant, the single principle which should guide the search for the categories is 'the clue to the discovery of the categories' (Kant 1978, A_{66}/B_{91}). Without such a clue, the search would be, he said, haphazard and unsystematic (Kant1978, A_{66}/B_{91}). Aristotle had made an attempt to enumerate the categories, but his attempt ended in failure because he did not proceed according to a single principle (Kant1978, A_{81}/B_{107}). He based his enquiry on empirical observations. But when an enquiry is carried on in this fashion, we can never be sure of its completion. We could never discover why just these concepts, and no others, have their seat in the understanding. Further, the concepts which we discover by empirical observations exhibit no order and systematic unity. Aristotle had based his enquiry on empirical observations with the result that he failed to offer a complete list of the categories. He omitted some fundamental and original concepts. Kant is referring here, we presume, to the category of causality and to the modal categories. Moreover, his list showed no order or a systematic unity. Kant attempted to avoid the defects of Aristotle's list of categories and decided to proceed according to a single principle. This principle, according to him, would necessarily be transcendental (Kant 1978, A_{67}/B_{92}) since the completeness of the table of categories could never be guaranteed by any principle other than transcendental. Kant made a twofold demand upon this principle. First, it must enable us to discover all the pure and original concepts of the understanding and secondly, it

should furnish an exact classification of them exhibiting their inner connexion in a system. This transcendental principle would act as the clue to the discovery of the categories.

Kant claimed that such a clue must be found by analysing the nature of the understanding since understanding, he asserted, was an absolute unity, self-contained and complete. The question that arises is - what was the principle or the transcendental clue that Kant followed?

The specific and clear formulation of the principle, which provided Kant with the transcendental clue to the discovery of the categories, is that to every form of judgment there corresponds a pure basic concept of the understanding. That this is what Kant meant is borne out by a passage where, after presenting his table of categories, he says: "This division is developed systematically from a common principle, namely, the faculty of judgment (which is the same as the faculty of thought) (Kant1978, $A_{80\text{-}81}/B_{106}$)."

Kant has recognized twelve forms of judgment under four heads, namely, quantity, quality, relation and modality. The three kinds of quantitative judgments are universal, particular and singular. The threefold classification of qualitative judgments, on the Kantian list, are affirmative, negative and infinite. The triad of relational judgments, according to Kant, are: categorical, hypothetical and disjunctive. Kant's classification of judgments with respect to modality is problematic, assertoric, and apodeictic. To these twelve forms of judgment correspond twelve categories. The twelve categories under the four heads quantity, quality, relation and modality are (1) unity, plurality and totality; (2) reality, negation and limitation; (3) of inherence and subsistence (substance and accident), causality and dependence (cause and effect), and community (reciprocity between agent and patient); and (4) possibility-impossibility, existence-nonexistence, necessity-contingency, respectively.

Kant has presented a threefold subdivision of categories under each of four heads. Kant in his Prolegomena said:

> … the third arises from the first and the second, joined in one concept (Kant 1950, p. 19).

Therefore, totality means only plurality considered as unity. However, what Kant called a predicable, as distinguished from a category, may also result from the combination of one category with another. Kant in the second edition of his critique (Kant 1978, B111) clearly said that the combination of the first and second categories yielding the third in each group requires a special act of understanding which is different from the act that is exercised in the case of the first or second category. Hence, it must not be supposed, that the third category is not primary. The point is more clearly stated in Kant's original reply in his letter to Johann Schultz. Kant said:

> For although the third category does certainly arise out of a uniting of the first and second, it does not arise out of their mere conjunction but rather out of a synthesis whose possibility itself constitutes a concept, and this concept is a particular category (Kant 1967, p. 111).

Kant presented arguments at two different stages in support of this principle serving as the transcendental clue: one, in the section titled *The Logical Employment of Understanding* in $A_{67\text{-}69}$/ $B_{92\text{-}94}$ and the other in the section titled *The Pure Concepts of the Understanding, or Categories* in $A_{76}/B_{102\text{-}105}$ of the *Critique*. The arguments in these two sections are of the highest importance for metaphysical deduction.

The argument at the first stage seeks to show that the basic function of understanding finds expression in and through the logical forms of judgment. The implication is left to be drawn that the basic concepts of the understanding, therefore, must be in accord with the logical forms of judgment.

The argument at the second stage refers the logical forms of judgment to the same operations of the understanding which are involved in the categorical synthesis of the intuitions. The argument thus serves to buttress up the view that the categories and the forms of judgment must be in accord with one another.

Kant's Argument in $_{A67\text{-}69}$ /$B_{92\text{-}94}$ of the Critique

The passage embodying Kant's argument to be considered here is too long to quote. Prichard who quoted the passage at length remarked: It is not worthwhile to go into all the difficulties of this confused and artificial passage (Prichard 1909, p. 146).

This is a harsh criticism; but it must be admitted that the passage requires the most patient examination if the obscurities are to be cleared. Fortunately for us, the obscurities have already been largely cleared by Paton. We may analyse Kant's argument in the passage into five steps following Paton's contention, as shown below:

1. Understanding is a power of knowing by means of concepts.
2. To know by means of concepts is to judge.
3. To judge is essentially to unite our ideas.
4. The different ways in which judgement unites our ideas are the forms of judgement… independently of the nature of the ideas themselves.
5. Consequently, the complete list of the forms of judgment is a complete list of the different ways in which understanding unites ideas by means of judgment; that is to say, it is a complete list of the functions of the Understanding (Paton 1936, p. 248)".

Kant's Argument in $A_{76}/B_{102\text{-}105}$ of the Critique

Kant's argument at the second stage may be briefly stated as follows: According to Kant, the general function of the understanding performs two types of unification. The 'function' which is referred to here is 'the work proper to understanding, namely, thinking or judging.' Kant disclosed these kindred operations of the understanding in the transition from his table of judgments to the table of categories in $B_{104/105}$ (Kant1978, $B_{104\text{-}105}$). The general function of thinking performs two kinds of unification at two different stages. It imposes unity on the

different ideas in a judgment as well as on the mere synthesis of different ideas in an intuition. Further, Kant elaborated his observation by saying that the understanding, in its use of concepts by means of analytic unity, brings into being the logical forms of judgment, and by means of synthetic unity of manifold of intuition in general, it introduces a transcendental content into its ideas. By disclosing these functions of the understanding, he has shown that there is an intimate relationship between the forms of thought and the pure concepts of the understanding.

The Functions of the Understanding and Types of Unity

Understanding Kant's argument at the second stage, necessitates a clarification of his view that the general function of the understanding performs two types of unification.

In thinking or judging we unite different ideas under one concept. Kant believed that all judgments are functions of unity in our ideas. Paton reckoned that what Kant really meant by saying that judgments unify ideas is that judgments unify our intuitions. In other words, while making judgments we hold different individuals before our minds by means of their common characteristics. These individuals 'are united in the sense that they are thought together in virtue of their common characteristics' (Paton 1936, p. 282). The concept of these common characteristics is considered to be the predicate of the judgment which comprehends, under it, all the individuals referred to by the subject-concept. In this way many possible cognitions are gathered into one. The whole judgment may be said to unify the individuals to whom it refers, and in the different forms of judgment the individuals referred to are united in various ways. This procedure of bringing different ideas under a concept has been described as analytic. The analytic aspect is present in all forms of judgment.

Thought more than unites different intuitions under a concept of their common characteristics; it also imposes unity on the 'mere synthesis' of various ideas in an intuition, whereby the given sense- impressions are combined into one intuition or one object. The 'mere synthesis' which is referred to here is the synthesis of imagination (Kant 1978, A_{78}/B_{104}). Imagination is treated by Kant as understanding working at a lower level.

Kant believed that knowledge arose from the joint operation of the sensibility and the understanding. According to him, Knowledge always refers to knowledge of an object. Now, if knowledge is to have objective reality, that is, to relate to an object, the object must be capable of being in some manner given. Kant said that it is only through the sensuous intuitions that objects are given to us. But objects as such are not given to us in intuition. Intuition always presents us with a manifold of the senses, or appearances. The manifold of senses, in order to be the object of possible experience, requires synthesis or unification. It is the task of the imagination to synthesise the manifold of given intuition. The knowledge which results from imaginative synthesis is crude and indistinct and it does not give us knowledge in the proper sense of the term. It requires analysis in order to get clear and distinct. This synthesis must be brought to concepts. This is considered to be the function of the understanding. Kant considered this conceptual synthesis to be

a necessary condition of all knowledge of objects. It is through this function of the understanding that we first obtain knowledge of an object The crude indistinct knowledge which results from imaginative synthesis is brought under a concept which originates from within the understanding itself through the process of analysis. According to Kant, concepts are rules or ways of synthesis or unification (Kant 1978, $A_{106)}$. For example, the concept of a triangle is a rule or direction about how to combine three straight lines to form a closed figure. Kant held that the concepts which originate from within the understanding itself and to which the given manifold of intuitions must conform in order to constitute one complex intuition of an object are called categories. These categories originate and are imposed by the nature of our thought itself, and not by the nature of our given sensations. We combine the *given* as a substance with different accidents, or as a ground which has certain consequences, and so on. According to Kant, every object of knowledge, besides the particular structure which we recognise by our empirical concepts, has a universal or categorical feature which is imposed by our thought. Thought is the ultimate source of unity of the synthesis of our intuition. We have seen that the general function of understanding, namely, thinking, imposes unity on the different ideas in a judgment. So, it can be said that the different forms of judgment which are the manifestations of the different ways of unification of ideas spring from the nature of the thought itself. On the other hand, thought by means of categories, which originate from within itself, unifies the mere synthesis of different ideas in an intuition. Thus, the forms of judgments and the categories both originate from the same source, namely, the function of the understanding. Therefore, Kant argued that the twelve logical forms of judgment provide the clue to the origin of the twelve corresponding a priori concepts or categories. Further, the understanding by means of synthetic unity introduces a 'transcendental content.' to the categories or to those ideas which originate from within itself. It is called 'transcendental' in the sense that it makes knowledge possible. The understanding itself introduces the formal content into each and every idea which originates from within itself by means of synthetic unity. Categories, we have seen, were regarded by Kant as the basic rules of synthesis. The given manifold, in order to be an intuition of an object of knowledge, must be united in accordance with a basic rule of synthesis. The way in which a category synthesises the given manifold is the form of that category. We sometimes, for example, synthesise the given manifold as a 'table' or a 'house.' We recognise the unity of the particular matter combined when we apply the empirical concept of 'table' or 'house'. But there are, according to Kant, certain ultimate principles or basic rules governing such empirical syntheses. These basic rules are imposed by our nature of thought itself, and not by the nature of our given sensation. Before applying the concept of 'house' or 'table' we must combine the given as a substance with different accidents. The way in which a category synthesises the given manifold constitutes the form of that category. Since it is a concept of synthesis, this formal content is introduced by the nature of thought by means of the function of synthesis. The formal content is called 'transcendental' in the sense that the categorical synthesis is the necessary condition of knowledge regarding an object.

Critique of an Objection to the Principle, act as Clue to the Discovery of the Categories

It is a widely accepted view that for Kant the forms of judgment are the forms of analytic judgments only. The view is so widely accepted that Paton is led to call it 'the orthodox view' or 'the orthodox theory' (Paton 1967, pp. 249, 268). This view has been made popular by Kemp Smith. He said, in his interpretation of Kant, that there are 'just as many categories as there are forms of the analytic judgment.' He insisted - 'This is how the principle of the metaphysical deduction must be interpreted' (Kemp Smith 1918, p. 183).

Now the most serious objection to Kant is that the categories, cannot in principle, be derived from the forms of judgment, because these forms, being the forms of the analytic judgments only, cannot furnish any clue to the categories which are principles of synthesis.

The orthodox view – namely, the view that for Kant, the forms of judgment are the forms of analytic judgments only - has been examined in detail and refuted decisively by Paton. The various arguments that have been presented by Paton in his different writings cannot and need not be repeated here. What we propose to do here is to emphasise the main points that go to show that orthodox view or interpretation is erroneous.

Kant has never said, as Kemp Smith would make us believe, that there are as many categories as there are forms of analytic judgments. While making a transition from the table of judgments to the table of categories, Kant said that '… there arise precisely the same number of pure concepts of the understanding which apply *a priori* to objects of intuition in general, as, in the preceding table, there have been found to be logical functions in all possible judgments. … These concepts we shall, with Aristotle, call *categories*, … (Kant 1978, A_{79}/B_{105}).' Evidently, Kant made it very clear in this passage that there are as many categories as there are logical functions (i.e., forms) in all possible judgments. The expression 'in all possible judgments' must be noted. It shows, without the least ambiguity, that the logical forms of judgments listed in Kant's table of judgments are the forms, not of analytic judgments only, but of all possible judgments and hence of synthetic judgments as well.

Kant said in the *Prolegomena* that a category is "… a concept of that synthetical unity of intuitions which can only be represented by a given logical function of judgments" (Kant 1950, p. 52). Paton, with reference to this passage, argued that a form of judgment cannot represent the synthetic unity of intuitions unless it is the form of synthetic as well as of analytic judgments (Paton 1967, p. 260). This shows that a form of judgment to which a category corresponds, according to Kant, cannot be the form of analytic judgments only. We may say, therefore, that the orthodox view championed by Kemp Smith is erroneous. The most serious objection to Kant's principle that act as the clue to the discovery of categories does not stand up to scrutiny.

Conclusion

We have discussed Kant's notion of categories at length with and how he derived his categories according to a single principle. As categories could not be defined without perpetrating a circle, he did not put forward any definition of categories. Instead of proposing a definition, he began with a kind of characterisation of categories in the *Critique.* We have examined Kant's own statement regarding categories that they are 'original pure concepts of synthesis that the understanding contains within itself a priori.' Kant also pointed out that there was a system in the procedure adopted by him in selecting those concepts.

According to Kant, categories are concepts of synthesis, because they represent, or correspond to, the basic ways in which certain (categorematic) concepts are united or synthesised into judgments. Therefore, Kant's categories are syncategorematic concepts. Moreover, they represent the ways in which the given indeterminate manifolds of sense are necessarily organised or synthesised into determinate objects of knowledge. Both the matter (i.e., the content) and the form of the categories are due to the understanding. This is why Kant said that the understanding contains them within itself. Categories are a priori, because they come to be known only through reflection on the ways in which any given concepts are used in judgments, and not by appealing to empirical observations. Kant's categories, on the other hand, are original concepts in the sense that they are underivative. Endowed with these qualifications, they are doubtless good enough candidates for categorical status. However, to prove that candidate categories are categories in the strong sense, it is necessary to show that they necessarily apply to all objects of knowledge. We are not concerned with this problem in the present paper. Kant has nevertheless shown that categories do apply to all objects of knowledge in his Transcendental Deduction and Analytic of Principles. As for the present work, we are concerned with pure concepts or categories of the understanding and not with schematised categories.

Kant's notion of categories would not be clear if we do not explain how those categories were derived by him. Kant's objective was to trace the pure concepts of the understanding or categories in the nature of the understanding according to a single principle in the metaphysical deduction of the categories in the *Critique*. Kant referred to the single principle which guides the search for the categories as 'the clue to the discovery of the categories.' The specific and clear formulation of the principle which served for Kant as the transcendental clue to the discovery of the categories is that to every form of judgment there corresponds a pure and basic concept of the understanding. We have noticed that the forms of judgments and the categories both originate from the same source, namely the function of the understanding, which is thinking. Accordingly, Kant argued that the twelve logical forms of judgments provided the clue to the origin of twelve corresponding a priori concepts or categories. Kant presented his arguments at two different stages in support of the principle that served as the transcendental clue. These arguments have been explained in detail in the present work.

A criticism against the principle which acts as the clue to the discovery of the categories has been discussed and an attempt has been made to answer the criticism.

References

Beck LW (1969) *Early German Philosophy: Kant and His Predecessors.* Cambridge, Massachusetts: Belknap.

Dryer D P (1966) *Kant's Solution for verification in Metaphysics*. London: George Allen and Unwin.

Hume D (1982) *An Enquiry Concerning Human Understanding*. Calcutta: Progressive Publishers.

Kant I (1950) *Prolegomena to any future metaphysics*. Translated by LW Beck. Indianapolis, New York: The Bobbs- Merrill Co. Inc.

Kant I (1967) Letter to Marcus Herz, February 1772. In *Kant: Philosophical Correspondence 1759-99*. Translated and Edited by A Zweig. Chicago: University of Chicago Press.

Kant I (1967) Letter to Johann Schultz, February17, 1784. In *Kant: Philosophical Correspondence 1759-99*. Translated and Edited by A Zweig. Chicago: University of Chicago Press.

Kant I (1968) On the Form and Principles of the sensible and Intelligible World, Known as "Inaugural Dissertation" In *Kant: Selected Pre-Critical Writings*, translated by GB Kerferd and De Walford. New York: Manchester University Press.

Kant I (1978) *Critique of Pure Reason*. Translated by NK Smith. London: The Macmillan Press Limited.

Kemp Smith N (1918) A Commentary to Kant's Critique of Pure Reason. London: Macmillan.

Locke John (1975) *An Essay concerning human understanding*. Edited by Peter H Nidditch. Oxford University Press.

Paton HJ (1936) *Kant's Metaphysic of Experience*. Vol-1. London: Macmillan.

Paton HJ (1967) The Key to Kant's Deduction of the Categories. In *Kant: Disputed Questions*. Edited by MS Gram. Chicago: Quadrangle Books.

Prichard HA (1909) *Kant's Theory of Knowledge*. Oxford: Clarendon.